



25th NATIONAL CONFERENCE ON COMMUNICATIONS (NCC) 2019

20-23 February 2019

Indian Institute of Science, Bangalore

[Message From
The General Co-Chairs](#)

[Message From
The TPC Co-Chairs](#)

[Conference
Committees](#)

[Plenary Talks](#)

[Tutorials](#)

[Industry Talks](#)

[Invited Talks](#)

[Technical Program](#)

[Search](#)

[Sponsors](#)

[Author Index](#)

[Contact/Map](#)

MESSAGE FROM THE GENERAL CO-CHAIRS

It is our pleasure to welcome you to the twenty fifth edition of the National Conference on Communications (NCC). NCC is now widely acknowledged to be a premier conference in India covering the fields of signal processing, communications, and networks. It is run by the Joint Telematics Group comprising Indian Institutes of Technology (IITs) and IISc.

NCC 2019 is very special since it is the Silver Jubilee edition of the conference. It is a great milestone for the conference to have run continuously and successfully for 25 years. This is a tribute to the vision of the faculty who founded this conference and to all those who conducted it all these years. To commemorate this historic occasion, we have put together a special "Silver Jubilee Reminiscences" event. It will be held on February 22 evening. A web page to commemorate this has also been set up – it features thoughts and comments from eminent researchers from India and abroad who have participated in NCC. Do look around the conference for surprise banners and other items related to this.

In this year's edition of NCC, we have come up with a program that blends high-quality contributed papers with invited talks, tutorials on cutting-edge topics by both academia and industry, and plenary talks by eminent researchers. NCC 2019 will feature three plenary talks, two industry keynotes, and five tutorials. Our goal has been to expose our participants from academia and industry to world-class, cutting-edge research and development.

We have also introduced a new Graduate Students Day event, in which several selected senior PhD students will present a short summary of their research work. This event will be held on February 21 afternoon. We hope that this will be appealing to students, faculty, and prospective employers. NCC 2019 will also feature cutting-edge demos from the 5G testbed project that is funded by the Department of Telecommunications and is a joint effort of several IITs, IISc, and SAMEER.

The plenary talks and industry keynotes will be held in the spacious and historic faculty hall in the main building of IISc. All other events will take place in the ECE department. We welcome you to check out our airy and bright new building in ECE. The recently inaugurated new lecture halls in this building will host several sessions and tutorials of NCC 2019. We hope you will enjoy the serene and green ambience of the ECE department. The conference proceedings will be available via download from the internet. The proceedings will also be accessible via IEEEXplore. In addition, three best paper awards and two best graduate student day talks have been instituted.

The success of NCC is due to the untiring efforts of its organizing committee and the army of volunteers from among the students and staff at IISc. We sincerely thank all the committee members for doing a wonderful job in putting together this major conference and ensuring that it runs smoothly. We thank the technical program committee for burning the midnight oil and putting together an exciting technical program.

Finally, we extend our thanks to the sponsors whose financial contributions made this conference possible. These sponsors will be exhibiting their cutting-edge products in their respective booths, and we encourage you to visit them.

We invite you to participate in and thoroughly enjoy NCC 2019.

Sincerely,

Neelesh B. Mehta and A. Chockalingam

GENERAL CO-CHAIRS

MESSAGE FROM THE TPC CO-CHAIRS

It gives us great pleasure to present to you the technical program for the silver jubilee edition of NCC. We hope that, like us, you will be impressed by the overall quality of the contributed papers this time around. A total of 224 papers were submitted for review, of which 90 papers made it to the final conference program, making the acceptance rate slightly over 40%. The relatively high acceptance rate is a reflection of the fact that the submissions were generally of a very high quality.

Roughly two-thirds (59/90) of the accepted papers were from the IITs/IISc, and the remaining one-third from other Indian institutes and universities (which includes the NITs and the IIITs). The relatively newer IITs of Bhubaneshwar, Dharwad, Gandhinagar, Hyderabad, Indore, Mandi and Patna accounted for 13 of the 59 IIT/IISc papers. While this is a commendable achievement, it is nonetheless sobering that, overall, half the accepted papers at NCC 2019 came from the established IITs (including Roorkee) and IISc.

For the first time in its history (to the best of our knowledge), NCC tried to remove any systemic bias in the review process by making it double-blind, meaning that besides the reviewers' identities not being revealed to the authors, the authors' identities were not made known to the reviewers as well. By our reckoning, this experiment was a success --- it was taken seriously and received broad support from authors and reviewers alike. We enthusiastically endorse repeating it in future editions of NCC. We would like to take this opportunity to thank the authors for largely adhering to the requirements of the double-blind review process. The reviewers were vigilant in flagging the handful of submitted papers that violated this policy. We would like to express our gratitude to the reviewers for their diligence, and more generally, for doing an excellent job with the review process overall.

Besides the contributed program, we have a wonderful set of plenary and invited talks. Prof. Ashutosh Sabharwal (Rice University), Dr. Yashwant Gupta (National Centre for Radio Astrophysics, Pune), and Prof. Ness Shroff (The Ohio State University) have consented to be the plenary speakers at NCC 2019. We have planned two industry keynote lectures (National Instruments and Qualcomm), and four sessions of invited talks to be given by faculty drawn from various institutions across India. The tutorial program on the first day of the conference is also excellent, with five tutorials on topics of current interest in communications, signal processing and machine learning being presented by internationally renowned experts.

A special feature of the technical program this time is Graduate Student Day, scheduled for the afternoon of Thursday, February 21. This is intended as a forum for soon-to-be-graduating PhD students to showcase their research work. We have selected 17 students from 13 different institutions to give short presentations on their work. You are requested to attend these sessions to provide encouragement to the future research stars in our field.

Yet another speciality in this year's NCC is the emphasis in 5G via dedicated tutorials from Rajeev Gangula on prototyping 4G/5G systems with OpenAirInterface, Aditya K. Jagannatham on mmWave radios, and Amod Anandkumar and Pallavi Kar on deep learning with Matlab, invited talks on 5G related topics from leading academics around the country, and several demo booths from startup companies in the areas of 5G. In addition, the booths from our financial sponsors are focussed on 5G related demos. We hope you will enjoy listening to these talks and visiting the demo booths.

We would like to take this opportunity to thank all the people who helped us put together the technical program. The plenary and invited sessions were organized with help from Neelesh Mehta and K.V.S. Hari. The contributed program is entirely due to the hard work of our TPC members, reviewers, and of course, the authors themselves. The Tutorials Committee of Prasanta Kumar Ghosh and Aditya Gopalan was responsible for organizing the tutorials. Graduate Student Day was the brainchild of Himanshu Tyagi, but it was realized through the efforts of Chandramani Singh and Kunal Chaudhury. The technical program booklet in your hands been put together beautifully by our Publications Chair, Soma Biswas. Timely updates to the website, particularly regarding the submission process and final program, were managed by Himanshu Tyagi. Finally, a big thanks to the general chairs Neelesh Mehta and A. Chockalingam, who ably oversaw the entire conference organization.

With these words, we welcome you to NCC 2019. We hope that you will enjoy the conference.

With best wishes,

Navin Kashyap, Dilip Krishnaswamy and Chandra Murthy

TPC CO-CHAIRS, NCC 2019

EXECUTIVE COMMITTEE

General Chairs	Neelesh B. Mehta (IISc Bangalore)
	A. Chockalingam (IISc Bangalore)
TPC Chairs	Chandra R Murthy (IISc Bangalore)
	Navin Kashyap (IISc Bangalore)
	Dilip Krishnaswamy (Reliance Jio)
Finance Chair	Rajiv Soundararajan (IISc Bangalore)
Publicity Chair	Shayan Srinivasa (IISc Bangalore)
Tutorial Chairs	Aditya Gopalan (IISc Bangalore)
	Prasanta Kumar Ghosh (IISc Bangalore)
Website Chair	Himanshu Tyagi (IISc Bangalore)
Publication Chair	Soma Biswas (IISc Bangalore)
Local Arrangements Chairs	Sriram Ganapathy (IISc Bangalore)
	Navin Kumar (Amrita School of Engineering)
	Sundeep Prabhakar Chepuri (IISc Bangalore)
Industry Liaison Chairs	Neelesh B. Mehta (IISc Bangalore)
	Dilip Krishnaswamy (Reliance Jio)
Graduate Student Day Chairs	Kunal Chaudhury (IISc Bangalore)
	Chandramani Singh (IISc Bangalore)

CONFERENCE COMMITTEES

Technical Program Committee

Cedric Adjih	INRIA	France
Samar Agnihotri	IIT Mandi	India
Shaik Rafi Ahamed	IIT Guwahati	India
M Jaleel Akhtar	IIT Kanpur	India
Sai Dhiraj Amuru	IIT Hyderabad	India
Gargeshwari Anand	IISc Bangalore	India
Sanya Anees	IIIT Guwahati	India
Kumar Appaiah	IIT Bombay	India
Dileep Aroor Dinesh	IIT Mandi	India
Ashwin Ashok	Georgia State University	USA
Sibi Raj B Pillai	IIT Bombay	India
Vineeth Bala Sukumaran	IIST Trivandrum	India
Jaiganesh Balakrishnan	Texas Instruments	USA
Adrish Banerjee	IIT Kanpur	India
Ramesh Battula	Malaviya NIT Jaipur	India
Bharath Bettagere	IIT Dharwad	India
Srikrishna Bhashyam	IIT Madras	India
Vimal Bhatia	IIT Indore	India
Ratnajit Bhattacharjee	IIT Guwahati	India
Arnav Bhavsar	IIT Mandi	India
Bharath Bhikkaji	IIT Madras	India
Prabir Kumar Biswas	IIT Kharagpur	India
Pravesh Biyani	IIIT Delhi	India
Vivek Bohara	IIIT Delhi	India
Santhana Krishnan Boopalan	University of Waterloo	Canada
Vivek Borkar	IIT Bombay	India

Manoj Bs	IIST	India
Amitalok Budkuley	The Chinese University of Hong Kong	Hong Kong
Saswat Chakrabarti	G. S. Sanyal School of Telecommunications	India
Sandip Chakraborty	IIT Kharagpur	India
Joydeep Chandra	IIT Patna	India
Sumohana Channappayya	IIT Hyderabad	India
Ajit Chaturvedi	IIT Kanpur	India
Sachin Chaudhari	IIIT Hyderabad	India
Nanda Kishore Chavali	Mathworks India Pvt Ltd.	India
Chandra Sekhar Chellu	IIT Madras	India
A. Chockalingam	IISc Bangalore	India
Ribhu Chopra	IIT Guwahati	India
Marceau Coupechoux	Telecom ParisTech	France
Govind D	Amrita Vishwa Vidyapeetham	India
Haresh Dagale	IISc Bangalore	India
Sam Darshi	IIT Ropar	India
Nabanita Das	ISI Kolkata	India
Raja Datta	IIT Kharagpur	India
Swades De	IIT Delhi	India
Kuntal Deka	IIT Goa	India
Bikash Dey	IIT Bombay	India
Kalpana Dhaka	IIT Guwahati	India

Sanjay Dhar Roy	NIT Durgapur	India
Harpreet Dhillon	Virginia Tech	USA
Debi Dogra	IIT Bhubaneswar	India
Ankit Dubey	NIT Goa	India
Santanu Dwari	Indian School of Mines	India
Jobin Francis	IIT Palakkad	India
Vikram Gadre	IIT Bombay	India
Ramakrishnan Ganesan	IISc Bangalore	India
Suryakanth Gangashetty	IIIT Hyderabad	India
Ravi Gangwar	IIT Dhanbad	India
Nithin George	IIT Gandhinagar	India
Debashis Ghosh	IIT Roorkee	India
Venkatesh Gopalakrishnan	IIT Madras	India
Ashutosh Gore	Qualcomm	India
Subrahmanyam Gorthi	IIT Tirupati	India
Siddhartan Govindasamy	F. W. Olin College of Engineering	USA
Abhishek Gupta	IIT Kanpur	India
Anubha Gupta	IIIT Delhi	India
Hari Gupta	IIT Varanasi	India
Rajiv Gupta	Terna Engineering College	India
Sanjeev Gurugopinath	PES University	India
J Harshan	IIT Delhi	India
Mohammad Hashmi	Nazarbayev University	Kazakhstan
Chinmay Hegde	Iowa State University	USA
Rajesh Hegde	IIT Kanpur	India

Ravi Hegde	IIT Gandhinagar	India
Ashraf Hossain	NIT Silchar	India
Yu-Chih Huang	National Taipei University	Taiwan
Tony Jacob	IIT Guwahati	India
Aditya Jagannatham	IIT Kanpur	India
Krishna Jagannathan	IIT Madras	India
Devendra Jalihal	IIT Madras	India
Thiruvengadam Jayaraman	Thiagarajar College of Engineering Madurai	India
Manasa K	IIT Hyderabad	India
Samudravijaya K	IIT Guwahati	India
Premkumar K.	IIITDM Kancheepuram	India
Sheetal Kalyani	IIT Madras	India
Vivek Kanhangad	IIT Indore	India
Nikhil Karamchandani	IIT Bombay	India
Veena Karjigi	SIT Tumkur	India
Gaurav Kasbekar	IIT Bombay	India
Navin Kashyap	IISc Bangalore	India
Salil Kashyap	IIT Guwahati	India
Mohammed Khan	IIT Hyderabad	India
Uday Khankhoje	IIT Madras	India
Nitin Khanna	IIT Gandhinagar	India
Manas Khatua	IIT Guwahati	India
Arzad Kherani	IIT Bhilai	India
Sri Rama Murty Kodukula	IIT Hyderabad	India
Mahesh Kolekar	IIT Patna	India
Sastry Komella	Naval Research Laboratory	USA

Prasad Krishnan	IIT Hyderabad	India
Dilip Krishnaswamy	IBM Research	India
Rakhesh Kshetrimayum	IIT Guwahati	India
Ankur Kulkarni	IIT Bombay	India
Abhinav Kumar	IIT Hyderabad	India
Animesh Kumar	IIT Bombay	India
Arun Kumar	IIT Delhi	India
Lalan Kumar	IIT Delhi	India
Pradeep Kumar	IIT Kanpur	India
Preetam Kumar	IIT Patna	India
Sudhir Kumar	IIT Patna	India
Brijesh Kumbhani	IIT Ropar	India
Chinmoy Kundu	Queen's University Belfast	United Kingdom (Great Britain)
Sumit Kundu	NIT Durgapur	India
Srinivas Kv	IIT BHU	India
V. Lalitha	IIIT Hyderabad	India
Ashok Kumar M.	IIT Palakkad	India
Srikanth Madikeri	Idiap Research Institute	Switzerland
Santi Maity	Indian Institute of Engineering Science and Technology Shibpur	India
Somnath Majhi	IIT Guwahati	India
Sudhan Majhi	IIT Patna	India
R Malmathanraj	NIT Trichy	India
D. Manjunath	IIT Bombay	India

Aashish Mathur	IIT Jodhpur	India
Neelesh Mehta	IISc Bangalore	India
Deepak Mishra	IIST	India
Rajiv Misra	IIT Patna	India
Parthajit Mohapatra	IIT Tirupati	India
Uma Mudenagudi	B. V Bhoomaraddi College of Engineering and Technology Hubli	India
Animesh Mukherjee	IIT Kharagpur	India
Biswajeet Mukherjee	IIITDM Jabalpur	India
Jayanta Mukherjee	IIT Kharagpur	India
Ravibabu Mulaveesala	IIT Ropar	India
Sriram Murali	Texas Instruments India	India
Chandra Murthy	IISc Bangalore	India
Hema Murthy	IIT Madras	India
Hemachandra N.	IIT Bombay	India
Kalpana Naidu	NIT Warangal	India
Vinayak Naik	BITS Pilani	India
Jayakrishnan Nair	IIT Bombay	India
R Nakkeeran	Pondicherry University	India
Lakshmi Natarajan	IIT Hyderabad	India
Karan Nathwani	IIT Jammu	India
V P Harigovindan	NIT Puducherry	India
Ram Bilas Pachori	IIT Indore	India
Ganapati Panda	IIT Bhubaneswar	India
Amritanshu Pandey	IIT BHU	India
Prem Pandey	IIT Bombay	India
Vinod Pankajakshan	IIT Roorkee	India

Aaqib Patel	IIT Hyderabad	India
Nagendra Pathak	IIT Roorkee	India
Moumita Patra	IIT Guwahati	India
Anil Prabhakar	IIT Madras	India
Vinod Prabhakaran	Tata Institute of Fundamental Research	India
Kmm Prabhu	IIT Madras	India
Pyari Pradhan	IIT Roorkee	India
Shankar Prakriya	IIT Delhi	India
Ranjitha Prasad	Tata Consultancy Services	India
Mahadeva Prasanna	IIT Dharwad	India
Niladri Puhan	IIT Bhubaneswar	India
A F M Sajidul Qadir	Samsung R&D Institute-Bangladesh	Germany
Manivasakan R	IIT Madras	India
Brijesh Rai	IIT Guwahati	India
Gaurav Raina	IIT Madras	India
B. Sundar Rajan	IISc Bangalore	India
Padmanabhan Rajan	IIT Mandi	India
Ketan Rajawat	IIT Kanpur	India
Alentattil Rajesh	IIT Guwahati	India
N s Rajput	IIT BHU	India
Ajit Rajwade	IIT Bombay	India
Swaminathan Ramabadran	Nanyang Technological University Singapore	Singapore
Venkatesh Ramaiyan	IIT Madras	India
Bhaskaran Raman	IIT Bombay	India

Shanmuganathan Raman	IIT Gandhinagar	India
Renu Rameshan	IIT Mandi	India
Barathram. Ramkumar	IIT Bhubaneswar	India
Preeti Rao	IIT Bombay	India
Shilpa Rao	IIIT Guwahati	India
Hemant Kumar Rath	Tata Consultancy Services	India
Mehul Raval	School of Engineering and Applied Science(SEAS), Ahmedabad University	India
Jithin Ravi	Universidad Carlos III de Madrid	Spain
Karun Rawat	IIT Roorkee	India
Meenakshi Rawat	IIT Roorkee	India
Vinay Ribeiro	IIT Delhi	India
Rajarshi Roy	IIT Kharagpur	India
Rajaram S	TCE Madurai	India
M Sabarimalai Manikandan	IIT Bhubaneswar	India
Seemanti Saha	NIT Patna	India
Jyotinder Sahambi	IIT Ropar	India
Pravas Ranjan Sahu	IIT Bhubaneswar	India
Rama Krishna Sai Gorthi	IIT Tirupati	India
Ravikant Saini	IIT Jammu	India
Amin Sakzad	Monash University	Australia
Nityananda Sarma	Tezpur University	India
Priyankoo Sarmah	IIT Guwahati	India
Pradeep Sarvepalli	IIT Madras	India

Vanlin Sathya	University of Chicago	USA
Jit Satyabrata	Banaras Hindu University	India
Mandha Damodaran Selvaraj	IITDM Kancheepuram	India
Debarati Sen	IIT Kharagpur	India
Shahid Shah	Islamic University of Science and Technology	India
Syed Shahnawazuddin	NIT Patna	India
Samar Shailendra	Tata Consultancy Services	India
Gvv Sharma	IIT Hyderabad	India
Prabhat Sharma	VNIT	India
Vinod Sharma	IISc Bangalore	India
Rahul Shrestha	IIT Mandi	India
Ajay Singh	IIT Jammu	India
Arun Kumar Singh	IIT Jodhpur	India
Chetna Singhal	IIT Kharagpur	India
Aditya Siripuram	IIT Hyderabad	India
Krishna Sivalingam	IIT Madras	India
Rajiv Soundararajan	IISc Bangalore	India
Babji Srinivasan	IIT Gandhinagar	India
Seshan Srirangarajan	IIT Delhi	India
Anand Srivastava	IIT Delhi	India
Suresh Sundaram	IIT Guwahati	India
Rajesh Sundaresan	IISc Bangalore	India
Arijit Sur	IIT Guwahati	India
Pravati Swain	NIT Goa	India
Anup Talukdar	Nokia Bell Labs	USA

Satyajit Thakor	IIT Mandi	India
Rahul Thakur	BITS Pilani Goa	India
Andrew Thangaraj	IIT Madras	India
G. Thavasi Raja	NIT Trichy	India
Lakshmi Narasimhan Theagarajan	IIT Palakkad	India
Madhan Thollabandi	Sterlite Tech	India
Anil Tiwari	IIT Jodhpur	India
Somanath Tripathy	IIT Patna	India
Prabhak Kumar Upadhyay	IIT Indore	India
Sreejith V	IIT Bhillai	India
Nidhin Vaidhiyan	Qualcomm India Pvt Ltd	India
Sundaram Vanka	Broadcom Corporation	USA
Manoj Varma	IISc Bangalore	India
Shailendra Varshney	IIT Kharagpur	India
Rahul Vaze	Tata Institute of Fundamental Research Mumbai	India
Mahendran Veeramani	IIT Tirupati	India
Rajbabu Velmurugan	IIT Bombay	India
Hrishikesh Venkataraman	IIIT Chittoor	India
Deepa Venkitesh	IIT Madras	India
Jiji Victor	College of Engineering Trivandrum	India
Saravanan Vijayakumaran	IIT Bombay	India
Sriram Vishwanath	University of Texas Austin	USA
Anil Kumar Vuppala	IIIT Hyderabad	India
Rajavaraprasad Yerra	IIT Hyderabad	India

CONFERENCE PLENARIES

P1: Rice RENEW: Empowering 3Rs of Research via Open Wireless Platforms & Testbeds

February 21 | 10:00-11:00

Faculty Hall, IISc main building



Ashutosh Sabharwal, Rice University, USA

Abstract:

Three R's are crucial for any scientific research endeavor that relies on an experimental component: repeatability, replicability, and reproducibility. In the first part of our talk, we reflect on our (personal) journey in developing and disseminating open-source wireless research platforms, and how they have shaped how we do research. Both successes and shortcomings from the past are at the heart of our next major step: the POWDER-RENEW platform, which is being designed to be the world's first open platform for next-generation massive-MIMO wireless research, with an emphasis to empower the 3R's for wireless research. In the second part of our talk, we reflect on our recent experimental research results in massive MIMO, to appreciate how flexible platforms can lead to novel research insights, and potentially shape the future wireless research and standards.

Biography:

Ashutosh Sabharwal works in two areas. His first area of research is wireless. He is the founder of WARP project (warp.rice.edu), an open-source project which is now in use at more than 125 research groups worldwide, and have been used by more than 450 research articles. He received 2017 Jack Neubauer Memorial and 2018 Advances in Communications Awards for full-duplex wireless. His second area of research is healthcare technologies. He is currently leading several NSF-funded center-scale projects, notably Rice RENEW (open-source massive MIMO) and "See below the skin" for non-invasive bio-imaging. He founded the Rice Scalable Health Labs which is developing a new engineering area called "bio-behavioral sensing." His research has led to four commercial spinoffs (one in wireless and three in healthcare).

CONFERENCE PLENARIES

P2: Signal processing challenges en route to understanding the Universe

February 22 | 9.30-10.30

Faculty Hall, IISc main building



Yashwant Gupta, TIFR, Pune

Abstract:

Contrary to what one might imagine, signal processing -- both algorithms and hardware -- play a crucial role in our quest to unravel the mysteries of the Universe. We will look at this interesting and challenging interplay between signal processing and astrophysics, primarily in the context of radio astronomy -- the branch of astronomy that works on the faint radio signals received from a host of natural phenomena (and maybe from extra-terrestrial intelligence!). From the complexity of processing of the weak signals from a multitude of receptor antennas to extract the signals of interest, to the algorithms that allow for combining of the signals to obtain useful images or high time resolution temporal data from astrophysical sources; from the challenges of real-time processing of the wide bandwidth signals, to the sophisticated off-line processing techniques that today span the realms of big data and machine learning : we will explore these various aspects, in the light of some of the existing modern radio observatories such as the Giant Metrewave Radio Telescope (GMRT) in India, as well as in the context of upcoming large international facilities of the future such as the Square Kilometre Array (SKA) project. Specific case studies to highlight the signal processing aspects will be presented.

Biography:

Professor Yashwant Gupta obtained his M.S. and Ph.D. in Radio Astronomy from the University of California, San Diego in 1990, after completing his Bachelor's degree in Electrical Engineering from IIT Kanpur in 1985. Since 1991, he has been working at the National Centre for Radio Astrophysics (NCRA, Pune) of the Tata Institute of Fundamental Research where he currently holds

the position of Senior Professor. For several years, he has been the Dean of the GMRT Observatory -- a world class instrument built and operated by the NCRA and located about 80 km from Pune. In March 2018, he took over as the Centre Director of NCRA.

His Ph.D. thesis was on the study of propagation of radio signals from pulsars through the inter-stellar medium of our Galaxy. His present research interests continue to be mainly in the area of pulsars : studying the details of their emission process and discovering new objects of this exotic species -- activities he continues to pursue actively using facilities such as the GMRT. He has over 100 papers in refereed journals and conference proceedings.

In addition to research in the astrophysics of pulsars, Prof Gupta also has significant interest and involvement in instrumentation and signal processing applications in radio astronomy. Since the early days of the GMRT, he has made significant contributions towards the development of the complex digital back-end receivers for the telescope. A considerable amount of his time at present also goes in leading the effort of a major technological upgrade of the GMRT, which is now almost complete. In addition, he also leads the technical involvement of India in the Square Kilometre Array (SKA) -- an international collaborative project to design and build the next generation global radio astronomy facility.

Prof Gupta was conferred the Shanti Swarup Bhatnagar Prize in the Physical Sciences, for the year 2007. He has been elected a fellow of the National Academy of Sciences of India, in 2007; and also elected a fellow of the Indian Academy of Sciences in January, 2008.

Prof Gupta has figured in several popular news and TV interviews, as well as some videos, many of which can be found on YouTube :

1. CNN-IBN program on "Superstars of Science" featuring Prof Gupta (Aug 2009) : "https://www.youtube.com/watch?v=_N1Ru5P39Yo"
2. Eureka with Yashwant Gupta (by Rajya Sabha TV) made in 2016 : "<https://www.youtube.com/watch?v=msMt22HJXDQ>"
3. Public talk by Prof Gupta at "Science at the Sabha" at Chennai, Feb 2017 : "<https://www.youtube.com/watch?v=7C3DyfDsrPE>"
4. Interview of Prof Gupta with Persistent Systems on work done for SKA (Oct 2018) : "https://wn.com/persistent_teams_with_india's_national_center_for_radio_astrophysics_to_develop_ska/video-details"

CONFERENCE PLENARIES

P3: A Fresh Look at an Old Problem: Network Utility Maximization—Convergence, Delay, and Complexity



February 23 | 10:00-11:00

Faculty Hall, IISc main building

Ness Shroff, Ohio State University, USA

Abstract:

Network Utility Maximization has been studied for resource allocation problems in communication networks for nearly two decades. Nonetheless, a major challenge that continues to remain open is how to develop a distributed congestion control and routing algorithm that can simultaneously provide utility optimality, fast convergence speed, and low delay. To address this challenge we take a fresh perspective on this old problem and develop a new algorithm that offers the fastest known convergence speed, vanishing utility optimality gap with finite queue length, and low routing complexity.

Our key contributions in this work are:

1. The design of a new joint congestion control and routing algorithm based on a type of inexact Uzawa method in the Alternating Directional Method of Multiplier
2. A new theoretical path to prove global and linear convergence rate without requiring the full rank assumption of the constraint matrix and
3. A clear path for implementing the proposed method in a fully distributed fashion.

Ness B. Shroff received his Ph.D. degree from Columbia University, NY in 1994 and joined Purdue university immediately thereafter as an Assistant Professor. At Purdue, he became Professor of the school of Electrical and Computer Engineering and director of a university wide center on wireless systems and applications (CWSA) in 2004. In July 2007, he joined the ECE and CSE departments at The Ohio State University, where he holds the Ohio Eminent Scholar Chaired Professorship of Networking and Communications. From 2009-2012, he also served as a Guest Chaired professor of Wireless Communications at Tsinghua University, Beijing, China, and currently holds an honorary Guest professor at Shanghai Jiaotong University in China and visiting position at the Indian Institute of Technology, Bombay.



Dr. Shroff's research interests span the areas of communication, networking, storage, cloud, recommender, social, and cyberphysical systems. He is especially interested in fundamental problems in learning, design, control, performance, pricing, and security of these complex systems. He currently serves as chair of the ACM MobiHoc steering committee, editor-at-large in the IEEE/ACM Trans. on Networking, and as senior editor of the IEEE Transactions on Control of Networked Systems. He also serves on the editorial boards of the IEEE Network Magazine, and the Network Science journal. He has served on the technical and executive committees of several major conferences and workshops. For example, he was the technical program co-chair of IEEE INFOCOM'03, the premier conference in communication networking, the technical program co-chair of ACM MobiHoc 2008, the General co-chair of WICON'08, and the conference chair of IEEE CCW'99. He has served as a keynote speaker and panelist on several major conferences in these fields. Dr. Shroff was also a co-organizer of the NSF workshop on Fundamental Research in Networking in 2003, and the NSF workshop on the Future of Wireless Networks in 2009.

Biography:

Dr. Shroff is a Fellow of the IEEE, and a National Science Foundation CAREER awardee. His papers have received numerous awards at top-tier venues. For example, he received the best paper award at IEEE INFOCOM 2006, IEEE INFOCOM 2008, and IEEE INFOCOM 2016, the best paper of the year in the journal of Communication and Networking (2005) and in Computer Networks (2003). He also received runner-up awards at IEEE INFOCOM 2005 and IEEE INFOCOM 2013. In addition, his papers have received the best student paper award (from all papers whose first author is a student) at IEEE WiOPT 2013, IEEE WiOPT 2012, and IEEE IWQoS 2006. Dr. Shroff is on the list of highly cited researchers from Thomson Reuters ISI (previously ISI web of Science) in 2014 and 2015, and in Thomson Reuters Book on The World's Most Influential Scientific Minds in 2014. He received the IEEE INFOCOM achievement award for seminal contributions to scheduling and resource allocation in wireless networks, in 2014.

TUTORIALS

MODERN METHODS OF MACHINE LEARNING



Sargur N. Srihari, University at Buffalo, USA

February 20

ECE Golden Jubilee Hall
Department of ECE

Abstract:

Artificial Intelligence (AI) methods of today are based on learning from examples--a methodology commonly known as machine learning. This tutorial provides an overview of modern methods of machine learning while introducing topics of current research interest—which include deep learning, probabilistic graphical models and reinforcement learning.

There will be four parts to the tutorial:

1. Overview of machine learning
2. Deep learning
3. Probabilistic graphical models
4. Reinforcement learning

We will begin with an overview of artificial intelligence and how simple machine learning techniques for classification and regression have largely replaced earlier knowledge-based methods. We will describe supervised/unsupervised learning, discriminative/generative models and specific algorithms such as logistic regression and SVMs.

The second part is about how deep learning differs from simple machine learning. We will describe commonly used deep learning architectures such as convolutional neural networks and recurrent neural networks. We will also describe several deep learning research topics, such as learning representations where variables are disentangled and decision making is simplified.

The third part will review probabilistic graphical models (PGMs). These are methods that allow reasoning with a diverse set of variables. They also allow explanations to accompany decisions. Methods of inference that overcome computational intractability will also be discussed.

The final part will concern reinforcement learning. Here the learning data is from the environment when an autonomous agent performs actions. We will describe methods known as Q-learning and policy learning. In particular deep reinforcement learning will be described.

Biography:

Srihari is a SUNY Distinguished Professor in the Department of Computer Science and Engineering at the University at Buffalo, The State University of New York. He held the Rukmini Govindachar chair in the School of Automation, Indian Institute of Science during 2018.

A laboratory that Srihari founded at Buffalo, known as CEDAR, developed the world's first automated system for reading handwritten postal addresses. It was deployed by the United States Postal Service-- which saved hundreds of millions of dollars in labor costs helping keep US postal rates lowest in the western world. A side-effect of this project was that it led to the task of recognizing handwritten digits to be considered the fruit-fly of AI methods.

Srihari also spent a decade developing AI and machine learning methods for forensics—focusing on pattern evidence such as latent prints, handwriting and footwear impressions. In particular, quantifying the value of handwriting evidence-- to allow presenting such testimony in US courts. Srihari has served on the National Academy of Sciences Committee on Identifying the Needs of the Forensic Science Community which led to an influential report. He has also served on NIJ-NIST committees on Human Factors in Fingerprint Analysis and Handwriting Comparison. At present he serves on the Houston Forensics Technical Advisory Board.

At Buffalo he teaches a sequence of three courses in artificial intelligence and machine learning:

- (i) Introduction to machine learning,
- (ii) Probabilistic graphical models and
- (iii) Deep learning. During 2018-19 he is teaching these subjects to over 400 graduate and undergraduate students.

Srihari's honors include: Fellow of the Institute of Electronics and Telecommunications Engineers (IETE, India) , Fellow of the IEEE, Fellow of the International Association for Pattern Recognition and distinguished alumnus of the Ohio State University College of Engineering. He received an Excellence in Graduate mentorship award from the University at Buffalo in 2018.

Srihari received a B.Sc. in Physics and Mathematics from the Bangalore University, a B.E. in Electrical Communication Engineering from the Indian Institute of Science and a Ph.D. in Computer and Information Science from the Ohio State University.

TUTORIALS

GENERATIVE TEXT-TO-SPEECH SYNTHESIS: FROM HMM-BASED SPEECH SYNTHESIS TO NEURAL END-TO-END TTS



Heiga Zen, Google Brain, Tokyo

February 20

MP Building Auditorium

Department of ECE

Abstract:

Generative model-based text-to-speech (TTS) synthesis has grown in popularity in the last a few years. Thanks to the recent progress of neural end-to-end approaches, it has reached the human-level quality in naturalness with flexibility to change its voice characteristics for synthesizing isolated sentences. Some of them have already been in production systems and served millions of queries. This tutorial will follow the progress of this technology from its fundamental concept to the real implementation, including the conventional statistical parametric speech synthesis to the latest neural end-to-end models.

Biography:

Dr. Heiga Zen, received his AE from Suzuka National College of Technology, Suzuka, Japan, in 1999, and PhD from the Nagoya Institute of Technology, Nagoya, Japan, in 2006. He was an Intern/Co-Op researcher at the IBM T.J. Watson Research Center, Yorktown Heights, NY, USA (2004–2005), and a Research Engineer at Toshiba Research Europe Ltd. Cambridge Research Laboratory, Cambridge, UK (2008–2011). At Google, he was with the Speech team from July 2011 to July 2018, then joined the Brain team from August 2018. Currently he is a Senior Staff Research Scientist at the Google Brain Tokyo team, Tokyo, Japan. His research interests include speech technology and machine learning. He was one of the original authors and the first maintainer of the HMM-based speech synthesis system (HTS).

TUTORIALS

OPEN-PROTOTYPING OF 4G/5G SYSTEMS WITH OPENAIRINTERFACE



Rajeev Gangula, EURECOM, France

February 20

MP Building Classroom

Department of ECE

Abstract:

This tutorial starts with a brief introduction to the OpenAirInterface Software Alliance (OSA) which aims at creating a global community developing open source software and hardware for 3GPP cellular networks. We then give an overview of the internals of the radio access network component of this project called openairinterface5g. We will describe the real-time processing architecture of OpenAirInterface (OAI) eNodeB (or basestation) with special emphasis on topics related to physical and access layers of the protocol stack. Finally, to show the usage of OAI in research, we present a scenario where an OAI enabled drone is used as a flying relay to enhance LTE connectivity to mobile users.

Biography:

Rajeev Gangula obtained his M.Tech degree from Indian Institute of Technology, Guwahati, in 2010, M.Sc. and Ph.D. degrees from Télécom ParisTech (Eurecom), France, in 2015. From October 2015 to December 2016 he was with Sequans communications, Paris, developing physical layer algorithms for LTE CAT-M chipsets. He is currently working at Eurecom as a research engineer building prototypes for autonomous aerial cellular relay drones capable of providing flexible and enhanced (LTE, 5G) connectivity to mobile users.

TUTORIALS

MMWAVE RADIO: THE NEXT ERA IN WIRELESS COMMUNICATION



Aditya K. Jagannatham, IIT, Kanpur

Abstract:

5G wireless networks are envisaged to support gigabit links to meet the ever-increasing demand for higher data rates. In this context, millimeter wave (mmWave) communication, which leverages the vast spectral opportunities in the mmWave band (30 – 300 GHz), has shown significant promise towards realizing the goals of next generation wireless systems. Further, due to the lower wavelength, multiple-input multiple-output (MIMO) is a natural technology of choice for such systems, which can significantly enhance their throughput. However, practical realization of mmWave MIMO technology is fraught with challenges since it is vulnerable to significantly higher losses arising from its higher carrier frequencies and also due to the increased hardware complexity for signal processing required to support the high bandwidth communication. To overcome these barriers, hybrid RF-baseband processing has emerged as a popular choice due to its lower complexity and improved RF/ baseband load distribution. This tutorial will present a brief analytical introduction to transceiver design and signal processing for mmWave MIMO systems, coupled with practical insights.

Biography:

Aditya K. Jagannatham received his Bachelors degree from the Indian Institute of Technology, Bombay and M.S. and Ph.D. degrees from the University of California, San Diego, U.S.A. From April '07 to May '09 he was employed as a senior wireless systems engineer at Qualcomm Inc., San Diego, California, where he was a part of the Qualcomm CDMA technologies (QCT) division. His research interests are in the area of next-generation wireless cellular and WiFi networks, with special emphasis on various 5G technologies such as massive MIMO, mmWave MIMO, FBMC, NOMA, Full Duplex and others. He is currently a Professor in the Electrical Engineering department at IIT Kanpur.

February 20

MP Building Auditorium

Department of ECE

TUTORIALS

INTRODUCTION TO DEEP LEARNING IN SIGNAL PROCESSING & COMMUNICATIONS WITH MATLAB



February 20

MP Building Classroom

Department of ECE

Dr. Amod Anandkumar, Ms. Pallavi Kar, Mathworks, India

Abstract:

Deep learning can achieve state-of-the-art accuracy for many tasks considered algorithmically unsolvable using traditional techniques. In this session, you can gain practical knowledge of the domain of deep learning and discover new MATLAB® features that simplify these tasks and eliminate the low-level programming. From prototype to production, you'll see demonstrations on training and deploying neural networks for signal processing and communications applications, including:

- Building deep learning models from scratch and performing transfer learning
- Understanding network behaviour using visualizations and other techniques
- Simplifying data labelling using MATLAB apps
- Accelerating deep learning training using multiple GPUs and computer clusters
- Generating code automatically from MATLAB algorithms and deploying it on enterprise applications and embedded devices

Bio: Dr. Amod Anandkumar is a senior team lead for signal processing and communications in the Application Engineering group at MathWorks India. Prior to this, he was a lead engineer with the Advanced Technology group at Samsung Research India, Noida where he developed physical layer techniques for LTE wireless communications systems and novel healthcare applications for smartphones. He was also a post-doctoral research fellow at the Biomedical Signal Analysis Lab, GE Global Research Bangalore, working on advanced beamforming techniques for ultrasound imaging and novel signal processing solutions for ICU patient monitoring systems, resulting in one US patent filing. Amod holds a B.Tech degree from National Institute of Technology Karnataka and a Ph.D. degree from Loughborough University, UK. His research interests include applied signal processing, next-generation wireless networks, computer vision, game theory, and convex optimization. He has published and reviewed papers in numerous international conferences and journals.

Ms. Pallavi Kar works as an application engineer at MathWorks in the area of Language of Technical Computing. She primarily focuses on the area of data analytics from intuition building and preprocessing of data to model development. Pallavi has five years of experience working across many industries. Over the years, she has worked on prognostics, lithium-ion batteries, model development and simulation, telematics, and server management. She has worked as a senior member of the Advanced Technologies team at Mahindra Reva Electric Vehicles in Bangalore. Pallavi holds a bachelor's degree in electronics and communication engineering and a master's degree in energy.

NCC 2019 GRADUATE STUDENT DAY WORKSHOP

The NCC 2019 Graduate Student Day workshop will be held at IISc Bangalore, India, as a part of NCC 2019. It will be a forum for PhD students to showcase their work in the areas of communication, networking and signal processing. The presenters will benefit from lively discussions and critical feedback on their work.

The following are the list of candidates that are shortlisted to present their work at the workshop:

Speaker	Thesis Title	Affiliation	Slot
Shikha Gupta	Deep CNN-based SMN Representation for Scene Recognition	IIT-Mandi	2:15-2:26 P.M.
Nazil Perveen	Spontaneous Facial Expression Recognition in Wild	IIT-Hyderabad	2:27-2:38 P.M.
Gowdham Prabhakar P G	Comparison of Ocular Parameters for Distraction Detection of Drivers in Cars	IISc Bangalore	2:39-2:50 P.M.
Sumit Datta	Efficient Compressed Sensing Magnetic Resonance Image Reconstruction for Clinical Applications	Tezpur University	2:51-3:02 P.M.
Anu Shaju Areeckal	Early Diagnosis of Osteoporosis Using Metacarpal Radiogrammetry and Texture Analysis	NIT-Suratkal	3:03-3:14 P.M.
Sweta Sharma	Large Scale Twin Support Vector Machine and its applications in Human Activity Recognition	South Asian University	3:15-3:26 P.M.
Rajendra Nagar	Multidimensional Reflection Symmetry: Theory, Algorithms, and Applications	IIT-Gandhinagar	3:27-3:38 P.M.
Tasleem Khan	Low Complexity Distributed Arithmetic based Pipelined VLSI Architectures for LMS Adaptive Filters	IIT-Guwahati	3:39-3:50 P.M.
Coffee Break	3:50-4:14 P.M.		
Praveen Jaraut	Digital Predistortion Linearization for Multi-band/Multi-channel Software Defined Transmitters	IIT-Roorkee	4:15-4:26 P.M.
Jinesh Jacob	Low Complexity Transmission in Few Mode Fibers using Limited Feedback of Principal Modes	IIT-Bombay	4:27-4:38 P.M.
Karan Gumber	Low Cost RF Predistortion for Carrier Aggregated Ultra-Wideband Signals	IIT-Roorkee	4:39-4:50 P.M.
Satish Kumar Tiwari	Estimation and Optimization of Design Parameters in Diffusive Molecular Nanonetworks	IIT-Indore	4:51-5:02 P.M.
Shaifu Gupta	Online Multivariate Resource Usage Prediction in Cloud Datacenters	IIT-Mandi	5:03-5:14 P.M.
Om Jee Pandey	Small World Models for Development of Wireless Sensor Network Services	IIT-Kanpur	5:15-5:26 P.M.
Priyanka Naik	libVNF: Library to build Virtual Network Functions	IIT-Bombay	5:27-5:38 P.M.
Rohit Kumar	Channel Selection in Dynamic Networks of Unknown Size	NIT-Delhi	5:39-5:50 P.M.
Vaibhav Kumar Gupta	Fair Subchannel Allocation Algorithms for the Inter Cell Interference Coordination with Fixed Transmit Power Problem	IIT-Bombay	5:51-6:02 P.M.

INDUSTRY KEYNOTE

**February 22 | 9.00 -9.30****Faculty Hall, IISc main building****MAKING 5G NR MMWAVE A COMMERCIAL REALITY****Dr. Kapil Bhattad**, Qualcomm, India**Abstract:**

5G NR (New Radio) is a new unified air interface for the next decade and beyond. It is designed to address a wide range of use cases covering enhanced mobile broadband, mission critical services and massive IoT, and to support new spectrum bands in both sub 6 GHz and mmwave. The talk will briefly cover the 5G NR foundational technologies that make this possible. It will then focus on the challenges in mmwave and mitigating techniques that make communication in mmWave possible on commercial handsets.

Bio: Kapil Bhattad received his B. Tech degree in Electrical Engineering from IIT Madras in 2002 and a Ph.D. degree in Electrical Engineering from Texas A&M University in 2007. He has been at Qualcomm research since then in San Diego from 2007 to 2011 and in Bangalore from 2011 onwards. He has contributed extensively to design, standardization, and commercialization of wireless communication technologies and chipsets covering 3G, 4G, 5G, IoT, and satellite communication systems. He has more than 200 patent applications pending and currently leads the 5G R&D effort at Qualcomm Wireless R&D at Bangalore.

INDUSTRY KEYNOTE



February 23 | 9.30 -10.00

Faculty Hall, IISc main building

STANDARDS-ORIENTED RESEARCH – NEED OF THE HOUR FOR INDIA

Ms. Pamela Kumar

Director General, TSDSI (Telecom Standards Development Society of India)
President & Founder Chair, CCICI (Cloud Computing Innovation Council of India)

Abstract:

The National Digital Communication Policy 2018 states :

At the current pace of digitisation and digitalisation, it is estimated that India's digital economy has the potential to reach one trillion USD by 2025.

With significant capabilities in both telecommunications and software, India, more than most countries, stands poised to benefit from harnessing new digital technologies and platforms to unlock productivity, thus catalysing economic growth and development

The 5G Vision for India has been defined as "5G technology has the potential for ushering a major societal transformation in India by enabling a rapid expansion of the role of information technology across manufacturing, educational, healthcare, agricultural, financial and social sectors. India must embrace this opportunity by deploying 5G networks early, efficiently, and pervasively, as well as emerge as a significant innovator and technology supplier at the global level."

The realization of these National Goals is dependent on the researchers & innovators in the country actively participating in steering the direction of emerging technologies to address the Unique challenges of India.

This talk will try to highlight the need for driving research to address India Specific requirements. This research can then be leveraged to develop standards, influence global standards and hence the direction of technology and product development.

TSDSI provides the one-stop platform :

To help researchers formulate research problems based on the Standardisation gaps needed for solving India's Unique challenges

To help researchers translate their research ideas and patents into Standards

The talk will provide the flow to achieve the above at TSDSI and invite the researchers at the conference to join hands in the development and adoption of the next generation of technology for India.

Bio: Since February 2017, Ms. Pamela Kumar has been appointed as the Director General of TSDSI. An alumnus of PEC Chandigarh, Rutgers University and IIM Bangalore, Pamela brings with her an experience of 30+ years in the communications, computers and semiconductor industry.

TSDSI is involved in mobilising the Technical experts in the country to develop standards based on National Priorities & requirements and influencing Global Standards, particularly at ITU, 3GPP and other forums. Major contributions are being driven to incorporate India's innovations to address Rural connectivity requirements for India and rest of the world. CCICI published a Framework & Roadmap for Cloud Computing Adoption in India and provided technical consultation on various aspects of the Cloud Computing Strategy for India. CCICI also formed the IoT for Smart Cities Task Force which published a series of reports related to Smart Cities Requirements Analysis, Smart Cities Reference Architecture, Smart Cities RFP Guidelines, Design & Planning of Smart Cities using IoT.

Prior to TSDSI, she has spent the first 10 years of her career at AT&T, Bell Labs in USA and at C-DOT in Bangalore. Later, she held leadership positions in Texas Instruments, IBM, Hewlett Packard Enterprise, R&D labs in India. She also did a short stint with 2 startups, setting up the R&D centers of Network Programs and Alliance Semiconductors. She holds 3 Patents granted by USPTO and has 5 patent applications pending in the Networking Accelerators domain. She has been a Keynote/ invited speaker in 60+ Local and Global forums.

Pamela is also the Founding Chair and current President of Cloud Computing Innovation Council of India (CCICI). She is on the Governing Board of Indian Financial Technologies and Allied Services (IFTAS) and PEC University of Technology. She is the Bharti Chair for IT Research and Member Research Council, UIET, Panjab University. She also mentors a few Startups in the ICT space.

Pamela is currently the Chair of the IEEE Charles Steinmetz Awards Committee. She was the first ever appointee from India in the IEEE Standards Association - Board of Governors, as the Member at large for 2011-12. She has held several other leadership roles as a senior member of IEEE - such as General Chair of IEEE ANTS Conference, Coordinator for IEEE Region 10 Industry Relations, Chair for IEEE Computer Society Chapter & Vice Chair of IEEE Bangalore section, etc.

INVITED TALKS

INV1: Invited Talks: Signal Processing

Room: MPA

Chair: Aditya Gopalan (IISc, India)

11:15 Tuning Free Algorithms for Sparse Estimation

Sheetal Kalyani (IIT Madras, India)

Sparse signal processing has attracted tremendous attention in the past two decades both from a theoretical point of view and in practical applications. However most of algorithms require prior knowledge of sparsity level/ outlier fraction or noise variance. In a practical scenario, the knowledge of such parameters cannot be assumed and estimating these parameter accurately is non-trivial. In this talk we discuss parameter free or tuning free variants of popular compressed sensing algorithms and also give theoretical guarantees for the same.

11:33 An Improved Multitasking Diffusion APA Based on Controlled Inter-Cluster Collaboration

Mrityunjoy Chakraborty (IIT, Kharagpur, India)

In distributed adaptive estimation, adaptive filters are placed at the nodes of a connected graph or network, which observe temporal data arising from different spatial sources with possibly different statistical profiles, and estimate certain parameter vector(s) of interest in a collaborative manner. Diffusion is a popular form of collaboration where each node shares its estimates with its neighbors, and also updates its estimates by using all the incoming estimate information from neighboring nodes. Depending on the number of parameter vectors to be estimated, adaptive networks can be classified into single-task and multitask networks. In a single-task network, all nodes collaboratively estimate a single optimal parameter vector, whereas in multitasking networks, the nodes are grouped into clusters and nodes within the same cluster are involved in estimating a common parameter vector. Different clusters generally have different (though related) tasks and the estimation is still carried out cooperatively as the data across the clusters may be correlated and, therefore, cooperation among clusters can be beneficial. This talk will focus on multitasking diffusion networks that achieve robustness against colored input by deploying adaptive filters at each node based on the well known affine projection algorithm (APA). It will be shown that while collaboration in principle is a useful step to enhance convergence performance, uncontrolled collaboration between clusters can at times lead to performance degradation, especially near steady state. To overcome this, we propose a controlled mode of inter-cluster collaboration via a suitable control variable which maintains collaboration in the right direction. The resulting improved multitask diffusion strategy exhibits faster convergence rate and lesser steady-state mean square deviation than the state-of-the-art. Simulation results in favor of the proposed scheme will also be presented.

INVITED TALKS

11:51 **Towards Self-Sustaining Devices Through Energy Harvesting**

Shankar Prakriya (IIT Delhi, India)

At present, the focus is on increasing lifetime of batteries in IoT type devices by utilising energy-efficient protocols. The ultimate goal however, is to make them self-sustaining. The QoS attained by links with energy harvesting nodes is limited greatly by the random nature of the energy harvested. In harvest-use (HU) architecture, we argue that carefully supplementing the harvested energy with as little battery energy as possible can dramatically increase performance and battery lifetime as well. We develop optimisation algorithms to maximise throughput or battery lifetime. It is shown that considerable savings in battery energy accrue when channel knowledge is available. Another approach to deal with the random variation in energy harvested is to utilise the harvest-store-use (HSU) architecture in which the harvested energy is stored in an energy buffer (rechargeable battery). Such nodes can become self-sustaining. We analyze performance of two-hop links with an energy harvesting relay, and use a continuous-state Markov chain to model energy in the buffer (instead of the discrete-state model that is commonly used). We compare performance with different buffer operation policies.

12:09 **Learning a Bandlimited Field From Samples Taken at Unknown-Locations on a Path**

Animesh Kumar (IIT Bombay, India)

Abstract TBD

12:27 **Sparse Sampling for Product Graphs and Tensors**

Sundeep Prabhakar Chepuri (IISc Bangalore, India)

In this talk, we consider the problem of subsampling and reconstruction of signals that reside on the vertices of a product graph, such as sensor network time series, genomic signals, or product ratings in a social network. Specifically, we leverage the product structure of the underlying domain and sample nodes from the graph factors. The proposed scheme is particularly useful for acquiring signals on large-scale product graphs and generalizes to acquiring multidomain signals, which can be represented using tensors with a known multilinear decomposition. The sampling sets are designed using a low-complexity greedy algorithm and can be proven to be near-optimal. Illustrations based on real data are provided for sampling 3D dynamic point clouds and for active learning in recommender systems. Joint work with Guillermo Ortiz-Jiménez, Mario Coutino, and Geert Leus.

INVITED TALKS

INV2: Invited Talks: 5G communications

Room: MPC

Chair: A. Chockalingam (IISc Bangalore, India)

13:45 Overview of the Indian 5G Testbed

Radha Krishna Ganti (IIT Madras, India)

The Department of Telecommunications has recently funded a pan-India 5G testbed. Eight institutes across India are involved in this project and the final goal is to build an end-to-end 5G testbed that can be used for research, testing and inter-operability. About 150 researchers with expertise in RF design, PCB design, FPGA board design, embedded system programming, networking and wireless systems are involved in this project. In addition, several Indian startups are also being roped in the project. In this talk, I will highlight the salient features, the end goals of the 5G testbed project. I will touch up on the technical specifications, the timelines and the key deliverables of the 5G testbed.

14:03 CeWiT's 5G Testbed Efforts

Babu Narayanan Koonampilli J (CEWIT, IIT Madras, India)

Test Beds play a critical role in the development of next generation technologies. They provide the platforms on which the novel ideas are experimented, implementation specifics are brought out and where the R&D team gets the first feel of the working of new technologies. It is a crucial stage between the simulation and the field trial phases. Test beds are also the learning grounds for research scholars and practicing engineers. The test beds continue to exist even after the technology arrives in the market - helping to stabilise, enhance and mature the technologies. 5G is the next generation cellular technology which is taking wireless communication in our everyday life to a different dimension. It has several advances in the radio technologies and revolutionary changes in the network architectures. Operators, Industry players and technology institutes in various countries have been developing 5G Test Beds to understand the technology implementation aspects better and assist building the 5G ecosystem. Not lagging behind, India is also in the track with several leading telecom companies and academic institutes building 5G Test beds in India, along with support from the Government of India. This talk will give an overall picture of the ingenuous end to end 5G test bed being developed in India collaboratively by eight technology institutes. It will also talk about how scholars can make use of the test bed when it is opened up.

14:21 Unified Control of Multi-RAT Radio Access Network: An SDN & NFV based Approach

Pranav Jha (IIT Bombay, India)

Telecom service providers are increasingly using a variety of Radio Access Technologies (RATs) for providing service to mobile subscribers. Although multiple RATs co-exist in today's wireless networks, each of these RATs is controlled independently, which may lead to sub-

INVITED TALKS

optimal utilization of resources in networks. An integrated control of the Multi-RAT Radio Access Network would help address this issue. We hereby propose a novel architecture to meet this requirement. The proposed architecture utilizes the concepts of Software Defined Networking (SDN) and Network Function Virtualization (NFV) to arrive at a unified control structure for Multi-RAT RAN. The architecture is also being formalized under the IEEE's standard development project P1930.1 as part of IEEE's Future Networks Initiative.

14:39 **Dynamic Mode Selection and Link Adaptation in 5G NR**

Sreenath R (Lekha Wireless, India)

Wireless channel, asynchronous information feedback, mobility, terrain are some of the underlying artifacts which can impact the link performance in a wireless network. In this talk, we briefly discuss the impact of these uncertainties on link performance and present some ideas and mechanisms to optimize mode selection and link adaptation in 5G NR. Seamless transitioning between Sub-6 GHz (FR1) and mmWave (FR2), automatic V2V / V2X connectivity on highways / urban grids are some of the potential cases which can benefit from discussed mechanisms.

14:57 **Fronthaul and Timing Standards for 5G**

Jishnu A (Tejas Networks, India)

Internet access is supposed to become pervasive like electricity with 5G. This would require a complete rearchitecture of how we build wireless infrastructure so that network can be implemented and operated in the most cost efficient way. The rearchitecture involves execution of most functionality in Central Office (CO) servers with ARM/x86 GPU enhanced with FPGA for hardware acceleration and moving the power amplification and part of PHY or higher layer into remote Radio Unit (RU). The link between this CO and RU, called Fronthaul brings in its own challenge like tighter latency and timing synchronization based on how this split is undertaken. In this talk, I will cover the various standards activity happening in Fronthaul and how the timing synchronization challenges are being addressed.

INV3: Invited Talks: Networks and Applications

Room: MPC

Chair: Parimal Parag (IISc Bangalore, India)

15:45 **Game-Theoretic Vaccination Against Networked SIS Epidemics and Impacts of Human Decision-Making**

Ashish Hota (IIT Kharagpur, India)

We study decentralized protection strategies against Susceptible-Infected-Susceptible (SIS) epidemics on networks. We consider a pop-

INVITED TALKS

ulation game framework where nodes choose whether or not to vaccinate themselves, and the epidemic risk is defined as the infection probability at the endemic state of the epidemic under a degree-based mean-field approximation. Motivated by studies in behavioral economics showing that humans perceive probabilities and risks in a nonlinear fashion, we specifically examine the impacts of such misperceptions on the Nash equilibrium protection strategies. We first establish the existence and uniqueness of a threshold equilibrium where nodes with degrees larger than a certain threshold vaccinate. When the vaccination cost is sufficiently high, we show that behavioral biases cause fewer players to vaccinate, and vice versa. We quantify this effect for a class of networks with power-law degree distributions by proving tight bounds on the ratio of equilibrium thresholds under behavioral and true perceptions of probabilities. We further characterize the socially optimal vaccination policy and investigate the inefficiency of Nash equilibrium.

16:03 **Making difficult things doable by leveraging communications: A case study of electric vehicles in India**

Ashok Jhunjhunwala (IIT Madras, India)

Abstract: The development of Communications has made a lot of difference to the world and its economy. We have seen Internet and Search emerge and WhatsApp, facebook and twitter rule the world. Electronic banking transactions, mobile payments is widespread driving e-commerce companies like Flipkart, Amazon and Snapdeal. OLA and Uber were unimaginable a couple of decades back. Swiggy, Zomato, Urban Clap and a whole lot of e-governance is riding on high-speed telecommunications. Internet of Things is and will continue to dominate our lives more and more. What one would see is that unimaginable things become possible. This would be the primary role of telecom as we go forward. It will impact totally unconnected areas. One such area is Electric Vehicles (EVs). India imports most of its oil and has fourteen out of the twenty most polluted cities in the world. The emergence of EVs would therefore be a god-send opportunity. The problem is that EVs are at present much more expensive as compared to petrol vehicles. Overseas, governments are providing up to 40% subsidy for EVs. Unfortunately India cannot afford that. If we wait, we will soon import EVs and its sub-systems, impacting a very strong industry, which contributes 7.1% of India's GDP. Can India make its EVs affordable today? It appears impossible. The paper presents an outline of what is being done today in India to optimise the battery resource, the key driver of costs for EVs. This would just not be possible without telecom and IoT.

16:21 **Optical Wireless Communication (OWC) Using LiFi Based System and Optical Camera Communication (OCC) Based System**

Shailesh Prabhu (Wipro, India)

I will share learnings from our experiments on Optical wireless communication (OWC) using LiFi based system and Optical Camera Communication (OCC) based system. I also will touch upon our point of view on the open areas that are potential candidates for research.

INVITED TALKS

16:39 **Millimeterwave Selection Optimization for Sustaining the 5G Use Cases**

Gourab Ghatak (IIIT Delhi, India)

Future wireless applications anticipate an explosion in the plethora of use-cases and services. Heterogeneous networks, comprising of multiple tiers of base stations (BSs) and operating in multiple radio access techniques (RATs) form an integral part of sustaining such diverse use-cases. One essential challenge of designing such complex networks is that of optimal user association to various tiers of BSs and to the available RATs. In the fourth generation (4G) networks, tier selection biasing is used mainly for load balancing. Offloading the user equipments (UEs) from the macro base station (MBS) to the small cell base stations (SBSs) is facilitated by a network-wide bias to expand the range of SBSs. On the contrary, in this talk, we discuss how to use various RAT-selection biases, each one associated to one of the fifth generation (5G) use-cases and designed to carefully satisfy the service requirements. In fact, to sustain the diverse use cases of 5G, a mobile operator will be able to define service-based logical partitions of its network over a common physical infrastructure. Network slicing facilitates the creation and management of such network instantiations (the network slices), each one composed by functions and parameters (e.g., the RAT bias in our work) tailored to address specific requirements [2]. In this talk, we focus on two specific RATs, i.e., sub-6GHz based LTE and mm-wave. We follow the specifications by 3GPP TS 23.501 [1], wherein, after the user, based on its service request, associates with one of the slices offered by the network, the bias value of the concerned slice is used for selection of the RAT in case the association is with an SBS. We will see how the diverse QoS requirements of different use-cases lead to a dramatically different RAT selectivity for the users.

16:57 **Distributed Smart Networks: A convergence of 5G, IoT, AI, and Blockchain**

Dilip Krishnaswamy (IBM Research, India)

This talk will discuss possibilities for distributed smart processing in emerging distributed virtualized network infrastructure. The talk will suggest how applications for smart cities and smart villages could leverage capabilities across a range of emerging technology areas such as 5G, IoT, AI, and Blockchain (post-hype), to deliver useful services to people.

INV4: Invited Talks: Communication theory and systems

Room: MPC

Chair: B. Sundar Rajan (IISc Bangalore, India)

11:45 **An OAI Based Testbed for Cellular-WiFi Convergence**

Bheemarjuna Reddy Tamma (IIT Hyderabad, India)

Cellular-Wi-Fi interworking architectures such as LTE Wi-Fi aggregation (LWA) and LTE Wi-Fi integration with IPsec tunnel (LWIP) are gaining momentum in the context of Multiple Radio Access Technology (Multi-RAT) in 5G. In this work, we present the design and implementation of LWA and LWIP prototypes which have been set up using OpenAirInterface (OAI) platform, and off-the-shelf Wi-Fi Access Point

INVITED TALKS

(AP) and smartphones. We then evaluate the performance of these prototypes with different Link Aggregation Strategies (LAS) for both UDP and TCP traffic. We have observed that, in a highly loaded Wi-Fi channel, when LWIP employs Wi-Fi only in Downlink (WoD) LAS, then sum of individual TCP flow throughput has improved by 28% as compared to LWIP operating with Flow Split (FS) LAS. We conclude the work with some interesting outcomes of various experiments which can be adopted as design principles for developing 5G Multi-RAT architectures.

12:03 **Solving a Distributed Stochastic Optimization Problem How Good is the Drift-Plus-Penalty Algorithm**

Bharath Bettagere (IIT Dharwad, India)

In this talk, I will present a distributed solution to the problem of minimizing the time average of a cost function subject to a set of constraints. These constraints are on the time averages of related stochastic processes called penalties. In such stochastic optimization problems, state of the system typically evolves in an independent and non-stationary fashion and the “common information” available at each node is distributed and delayed. This framework forms an integral part of many important problems in wireless networks and computer science such as scheduling, routing, resource allocation and crowd sensing, to name a few. In this talk, I will propose an approximate distributed Drift-Plus-Penalty (DPP) algorithm, and show that it achieves a time average cost (and penalties) that is within a “small” constant of the optimal cost (and constraints) with high probability. More importantly, I will show that the proposed algorithm converges almost surely to the optimal solution. I will also present an application in wireless sensor network to corroborate some of our theoretical findings through simulation results.

12:21 **Not Just Age but Age and Quality of Information**

Rahul Vaze (TIFR Mumbai, India)

A versatile scheduling problem to model a three-way tradeoff between delay/age, distortion, and energy is considered. The considered problem called the age and quality of information (AQI) is to select which packets to transmit at each time slot to minimize a linear combination of the distortion cost, the age/delay cost and the energy transmission cost in an online fashion. AQI generalizes multiple important problems such as age of information (AoI), the remote estimation problem with sampling constraint that specializes to AoI, the classical speed scaling problem among others. The worst case input model is considered, where the performance metric is the competitive ratio. A greedy algorithm is proposed that is shown to be 2-competitive, and independent of all parameters of the problem.

12:39 **Authenticated Communication over Multiple Access Channels with Adversarial Users**

Bikash K Dey (IIT Bombay, India)

We study authenticated communication over two- user multiple access channels (MAC) where one of the users is possibly adversarial. When both users behave non-adversarially, we want their messages to be decoded reliably. However, we also want to ensure that an adversarial user cannot cause an undetected error on the other (honest) user’s message. We show that the following three-phase scheme is rate-optimal: a standard MAC code is first used to achieve unauthenticated communication; this is followed by two authentication

phases where each user authenticates their message treating the other user as a possible adversary. We show that the authentication phases can be very short since this form of authentication itself, when possible, can be achieved for message sets whose size grow doubly exponentially in blocklength. This leads to our result that the authenticated communication capacity region of a discrete memoryless MAC is either zero or the (unauthenticated) MAC capacity region itself. This also, arguably, explains the similar nature of authenticated communication capacity of a discrete memoryless point-to-point adversarial channel recently found by Kosut and Kliewer (ITW, 2018). We also obtain analogous results for additive Gaussian noise channels.

12:57 **Adversarial Attacks on Next-Generation Wireless Networks by Cognitive Radios**

Harshan Jagadeesh (IIT Delhi, India)

Next-generation wireless networks are likely to include heterogeneous composition of numerous devices such as routers, antenna-arrays, leaky coaxial cables, UAVs, RFID tags etc., which will coherently operate to carry out specific objectives. Typical applications of such networks include (i) cyber physical systems such as urban transportation and smart-grids, (ii) ad-hoc networks such as wireless sensor networks, vehicular networks, and (iii) cellular networks such as conventional as well as UAV-based networks. With potential application to the above networks, we first discuss new threat models on wireless security arising out of a cognitive attacker which can instantaneously manipulate the transmitted symbols in the air, akin to man-in-the-middle attacks. Subsequently, we discuss suitable attack-detection strategies and countermeasures using tools inspired by physical-layer as well as cross-layer ideas.

TECHNICAL PROGRAM

BP1: Best Paper Awards Session 1

Thursday, February 21 11:45 - 1:15

Venue: GJH

Chair: Avhishek Chatterjee (IIT Madras, India)

Channels with Action Dependent States and Common Reconstructions

11:45- 12:03

Viswanathan Ramachandran (IIT Bombay, India)

Sibi Raj B Pillai (IIT Bombay, India)

Vinod M Prabhakaran (TIFR, India)

In channels with action dependent states, a common message is conveyed using two encoders operating sequentially, viz. an action encoder and a channel encoder. The actions drive the output of a discrete-memoryless channel (DMC), which in turn forms the state process for the DMC between the channel encoder and receiver. Assuming non-causal knowledge of the state-process at the channel encoder, a single letter characterization of the capacity is known in the discrete memoryless case. We consider the action dependent state channel with a common message and an additional private message at the channel encoder, along with common reconstructions (CR) of the state process at the channel encoder and the decoder. Capacity characterizations for the discrete memoryless and Gaussian versions are presented. As a consequence, we settle the capacity characterization of the Gaussian action dependent channel with only a common message and CR. Moreover, by identifying a connection to degraded message sets multiple access channel (deg-MAC) models studied in literature, we establish the capacity regions for the discrete and Gaussian versions of deg-MAC with CR constraint.

Conditions for Optimality of Superposition Coding in Discrete Memoryless Broadcast Channels

12:03 - 12:21

Harikumar Krishnamurthy (IIT Madras, India)

Parikshit Hegde (IIT Madras, India)

Andrew Thangaraj (IIT Madras, India)

The capacity region of general discrete-memoryless broadcast channels (DMBCs) with two receivers is an open problem of considerable research interest. The optimality of superposition coding in three specific cases of the DMBC is considered. For a DMBC with binary input, symmetric output and output cardinality at most 3, superposition coding is shown to be optimal. For equal-capacity DMBCs with any input cardinality, superposition coding is shown to be suboptimal if each channel has a capacity-achieving input distribution that is not capacity-achieving for the other channel. For an equal-capacity DMBC with binary input, superposition coding is shown to be optimal if and only if the two channels are more-capable comparable even without output symmetry. These results improve upon the previously known conditions for optimality of superposition coding in DMBCs.

TECHNICAL PROGRAM

Maximally Recoverable Codes with Hierarchical Locality

12:21 – 12:39

Aaditya M Nair (IIIT Hyderabad, India)

V. Lalitha (IIIT Hyderabad, India)

Maximally recoverable codes are a class of codes which recover from all potentially recoverable erasure patterns given the locality constraints of the code. In earlier works, these codes have been studied in the context of codes with locality. The notion of locality has been extended to hierarchical locality, which allows for locality to gradually increase in levels with the increase in the number of erasures. We consider the locality constraints imposed by codes with two-level hierarchical locality and define maximally recoverable codes with data-local and local hierarchical locality. We derive certain properties related to their punctured codes and minimum distance. We give a procedure to construct hierarchical data-local MRCs from hierarchical local MRCs. We provide a construction of hierarchical local MRCs for all parameters. For the case of one global parity, we provide a different construction of hierarchical local MRC over a lower field size.

Probability Mass Functions for which Sources have the Maximum Minimum Expected Length

12:39 – 12:57

Shivkumar K Manickam (IISc Bangalore, India)

Let P_n be the set of all probability mass functions (PMFs) (p_1, p_2, \dots, p_n) that satisfy $p_i > 0$ for $1 \leq i \leq n$. Define the minimum expected length function $L_D : P_n \rightarrow \mathbb{R}$ such that $L_D(P)$ is the minimum expected length of a prefix code, formed out of an alphabet of size D , for the discrete memoryless source having P as its source distribution. It is well-known that the function L_D attains its maximum value at the uniform distribution. Further, when n is of the form D^m , with m being a positive integer, PMFs other than the uniform distribution at which L_D attains its maximum value are known. However, a complete characterization of all such PMFs at which the minimum expected length function attains its maximum value has not been done so far. This is done in this paper.

Achieving Secrecy Capacity of Minimum Storage Regenerating Codes for all Feasible (n, k, d) Parameter Values

12:57 - 13:15

V. Arvind Rameshwar (IISc Bangalore, India)

Navin Kashyap (IISc Bangalore, India)

This paper addresses the problem of constructing secure exact-repair regenerating codes at the MSR point for all feasible values of the parameters. The setting involves a passive eavesdropper who is allowed to observe the stored contents of, and the downloads into, an l -subset of the n nodes of a distributed storage system (DSS). The objective is to achieve perfect secrecy between the eavesdropped symbols and the file stored on the DSS. Previous secure code constructions (most notably that by Rawat et al.) tackle the problem only for the restricted case wherein the number, d , of helper nodes aiding in the recovery of a failed node is equal to $n-1$. This paper builds on Rawat's work, by combining Gabidulin pre-coding and an MSR construction by Ye and Barg to prove the achievability of secrecy capacity at the MSR point for all allowed values of d .

TECHNICAL PROGRAM

SP1: Speech Recognition

Thursday, February 21 11:45 – 12:57

Venue: MPA

Chair: Gayadhar Pradhan (NIT Patna, India)

Data-pooling and multi-task learning for enhanced performance of speech recognition systems in multiple low resourced languages

11:45- 12:03

Madhavaraj A (IISc Bangalore, India)

A G Ramakrishnan (IISc Bangalore, India)

We present two approaches to improve the performance of automatic speech recognition (ASR) systems for GujaratiTamil and Telugu. In the first approach using data-pooling with phone mapping (DPPM), a deep neural network (DNN) is trained to predict the senones for the target language; then we use the feature vectors and their alignments from other source languages to map the phones from the source to the target language. The lexicons of the source languages are then modified using this phone mapping and an ASR system for the target language is trained using both the target and the modified source data. This DPPM approach gives relative improvements in word error rates (WER) of 5.1% for Gujarati, 3.1% for Tamil and 3.4% for Telugu, over the corresponding baseline figures. In the second approach using multi-task DNN (MT-DNN) modeling, we use feature vectors from all the languages and train a DNN with three output layers, each predicting the senones of one of the languages. Objective functions of the output layers are modified such that during training, only those DNN layers responsible for predicting the senones of a language are updated, if the feature vector belongs to that language. This MT-DNN approach achieves relative improvements in WER of 5.7%, 3.3% and 5.2% for Gujarati, Tamil and Telugu, respectively.

On the Role of Linear, Mel and Inverse-Mel Filter bank in the Context of Automatic Speech Recognition

12:03 - 12:21

Hemant K. Kathania(NIT Sikkim, India)

Syed Shahnawazuddin (NIT Patna, India)

Waqar Ahmad (NIT Calicut, India)

Nagaraj Adiga (University of Crete Greece)

In the context of automatic speech recognition (ASR), the power spectrum is generally warped to the Mel-scale during front-end speech parameterization. This is motivated by the fact that, human perception of sound is nonlinear. The Melfilterbank provide better resolution for low-frequency contents while a greater degree of averaging happens in the high-frequency range. The work presented in this paper aims at studying the role of linear, Mel and inverse-Mel filterbanks in the context of speech recognition. It is well known that, when speech data is from high-pitched speakers like children, there is a significant amount of relevant information in the high-frequency region. Hence, down-sampling the information in that range through Mel-filterbank reduces the recognition performance. On the other hand, employing inverse-Mel or linear-filterbanks are expected to be more effective in such cases. The same has been experimentally validated in this work. To do so, an ASR system is developed on adults' speech and tested using data from adult as well as child speakers. Significantly improved recognition rates are noted for children's as well adult females' speech when linear or inverseMel filterbank is used. The use of linear filters results in a relative improvement of 21% over the baseline.

TECHNICAL PROGRAM

Speaking-Rate Adaptation of Automatic Speech Recognition System through Fuzzy Classification based Time-Scale Modification

12:21 – 12:39

S. Shahnawazuddin (NIT Patna, India)
Hemant K. Kathania (NIT Sikkim, India)
Nagaraj Adiga (University of Crete Greece)
B. Tarun Sai (NIT Patna, India)
Waquar Ahmad (NIT Calicut, India)

In this paper, we study the role of speaking-rate adaptation (SRA) of automatic speech recognition (ASR) systems. The performance of an ASR system is reported to degrade when the speaking-rate is either too fast or too slow. In order to simulate such a situation, an ASR system was trained on adults' speech and used for transcribing speech data from adult as well as child speakers. Earlier studies have shown that, speaking-rate is significantly lower in the case of children when compared to adults. Consequently, the recognition performance for children's speech was noted to be very poor in contrast to adults' speech. To improve the recognition performance with respect to children's speech, speaking-rate was explicitly changed using time-scale modification (TSM). A recently proposed TSM approach based on fuzzy classification of spectral bins has been explored in this regard. The fuzzy-classification-based TSM technique is reported to be superior to state-of-the-art approaches. Effectiveness of the said TSM technique has not been studied yet in the context of ASR. The experimental studies presented in this paper show that SRA based on fuzzy classification results in a relative improvement of 30% over the baseline.

Instantaneous Frequency Features for Noise Robust Speech Recognition

12:39 – 12:57

Shekhar Nayak (IIT Hyderabad, India)
Shashank Dhar B. (Georgia Institute of Technology, Atlanta, USA)
Saurabhchand Bhati (IIT Hyderabad, India)
Koilakuntla Bramhendra (IIT Hyderabad, India)
K. Sri Rama Murty (IIT Hyderabad, India)

Analytic phase of the speech signal plays an important role in human speech perception, specially in the presence of noise. Generally, phase information is ignored in most of the recent speech recognition systems. In this paper, we illustrate the importance of analytic phase of the speech signal for noise robust automatic speech recognition. To avoid phase wrapping problem involved in the computation of analytic phase, features are extracted from instantaneous frequency (IF) which is time derivative of analytic phase. Deep neural network (DNN) based acoustic models are trained on clean speech using features extracted from the IF of speech signals. Robustness of IF features in combination with mel-frequency cepstral coefficients (MFCCs) was evaluated in varied noisy conditions. System combination using minimum Bayes risk decoding of IF features with MFCCs delivered absolute improvements of upto 13% over MFCC features alone for DNN based systems under noisy conditions. The impact of the system combination of magnitude and phase based features on different phonetic classes was studied under noisy conditions and was found to model both voiced and unvoiced phonetic classes efficiently.

TECHNICAL PROGRAM

RF1: Antenna Design

Thursday, February 21 11:45 - 13:15

Venue: MPC

Chair: Meenakshi Rawat (IIT Roorkee, India)

Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics

11:45 – 12:03

Ameya Kadam (DJSCE, India)

Amit Deshmukh (DJSCE, India)

Akshay Doshi (DJSCE, India)

Sanjay Deshmukh (Mumbai University & DJSCE, India)

Kamala Prasan Ray (DIAT, Pune)

This paper proposed a planar, low cost, simple, and compact printed microstrip-fed circular monopole ultrawideband antennas with band-notched characteristics. By introducing pair of open slits and varying angular separation between slits on the circular patch the band notched characteristics can be obtained. The proposed antennas are successfully simulated, designed, fabricated on FR-4 substrate. The measured results show that the proposed antenna with dimensions of $65\text{mm} \times 65\text{ mm} \times 1.6\text{mm}$ has a bandwidth over the frequency band 2.00-10.6 GHz with $\text{VSWR} \leq 2$, except 2.5-3.95GHz with circular antenna having pair of slits. The presented antennas shows nearly omnidirectional radiation pattern, stable gain, small group delay variation at working frequencies. Satisfactory results have been obtained in frequency and time-domain analysis of the proposed structure.

Analysis and Resonant Length Formulation of Dual Band Microstrip Antenna with Modified Ground

12:03 – 12:21

Poonam Kadam (DJSCE, India)

Amit Deshmukh (DJSCE, India)

Akshay Doshi (DJSCE, India)

Sanjay Deshmukh (Mumbai University & DJSCE, India)

A dual band microstrip antenna with defected ground plane is proposed in this paper. The dual bands are achieved by appropriately embedding pair of slots on the ground plane to tune TM30 mode resonant frequency. These slots essentially reduces TM30 mode frequency and places it closer to fundamental mode frequency. Broadside radiation pattern is observed at both the frequencies with maximum gain of nearly 3 dBi. The effect of the ground slot is studied and are portrayed by showing resonance plot for variation in the slot length. Empirical formulation of resonant length for fundamental as well as higher order TM30 mode frequency is suggested in the paper. The calculated frequencies obtained using the proposed formula closely matches with the simulated values. Thus these formulas can be applied to design dual band antenna with similar configuration at any given operating frequency

TECHNICAL PROGRAM

Proximity Fed Broadband Equilateral Triangular Microstrip Antenna Using Parasitic Rectangular Patches

12:21 – 12:39

Sanjay Deshmukh (Mumbai University & DJSCE, India)

Amit Deshmukh (DJSCE, India)

Akshay Doshi (DJSCE, India)

Gap coupled design of proximity fed equilateral triangular microstrip antenna along with variations of rectangular microstrip antennas are proposed. Design of equilateral triangular microstrip antenna provides a bandwidth of 302 MHz (25%) with broadside gain of 7.6 dBi. Enhancement in bandwidth and gain is realized by gap coupling rectangular patches. Optimum result with VSWR bandwidth of more than 670 MHz (57%), showing peak gain of 9.7 dBi is obtained in gap coupled configuration with two adjacent layers of two and four rectangular patches. Proposed design is simpler in implementation and measured results show close agreement.

Slot Cut Modified Triangular Shape Microstrip Antenna for Circular Polarization

12:39 – 12:57

Akshay Doshi (DJSCE, India)

Amit Deshmukh (DJSCE, India)

Sanjay Deshmukh (Mumbai University & DJSCE, India)

Kamala Prasan Ray (DIAT, Pune)

Modified design of triangular microstrip antenna obtained by combining two triangular shape patches is discussed. The offset distance between two patches tunes the spacing between orthogonal mode frequencies of the patch which yields circular polarized response with axial ratio bandwidth of 5%. Further to realize compact configuration, slot cut design of offset triangular overlap patches is studied. The rectangular slot helps in tuning the spacing between offset triangular patches along with offset distance to yield circular polarization. The detailed parametric study for the effects of slot along with offset distance in circular polarized antenna is presented. The tuning of orthogonal frequencies is obtained at lower offset distance between two patches for slot cut antenna. The optimum axial ratio bandwidth of 70 MHz in 1500 MHz frequency band is obtained along with broadside gain of 7 dBi and VSWR Bandwidth of 747 MHz. In optimum case slot cut overlapped design offers 42% reduction in patch area.

A Circular Fractal Antenna Array

12:57 - 13:15

Rahul Chauhan (DAIICIT, India)

Sanjeev Gupta (DAIICIT, India)

Modified design of triangular microstrip antenna obtained by combining two triangular shape patches is discussed. The offset distance between two patches tunes the spacing between orthogonal mode frequencies of the patch which yields circular polarized response with axial ratio bandwidth of 5%. Further to realize compact configuration, slot cut design of offset triangular overlap patches is studied. The rectangular slot helps in tuning the spacing between offset triangular patches along with offset distance to yield circular polarization. The detailed parametric study for the effects of slot along with offset distance in circular polarized antenna is presented. The tuning of orthogonal frequencies is obtained at lower offset distance between two patches for slot cut antenna. The optimum axial ratio bandwidth of 70 MHz in 1500 MHz frequency band is obtained along with broadside gain of 7 dBi and VSWR Bandwidth of 747 MHz. In optimum case slot cut overlapped design offers 42% reduction in patch area.

TECHNICAL PROGRAM

BP2: Best Paper Awards Session 2

Thursday, February 21, 14:15 - 15:45

Venue: GJH

High Performance Multiplierless Serial Pipelined VLSI Architecture for Real-Valued FFT

14:15- 14:33

Jinti Hazarika (IIT Hyderabad, India)

Mohd Tasleem Khan (IIT Guwahati, India)

Shaik Rafi Ahamed (IIT Guwahati, India)

Harshal B. Nemade (IIT Hyderabad, India)

This paper presents a high-performance multiplierless serial pipelined architecture for real-valued fast Fourier transform (FFT). A new data mapping scheme (DMS) is suggested for the proposed serial pipelined FFT architecture. The performance is enhanced by performing FFT computations in $\log_2 N - 1$ stages followed by a select-store-feedback (SSF) stage, where N is the number of points in FFT. Further enhancement in performance is achieved by employing quarter-complex multiplierless unit made up of memory and combinational logic in every stage. The memory stores half number of partial products while the remaining partial products are taken care by external combinational logic. Compared with the best existing scheme, the proposed design reduces the computational workload on half-butterfly (H-BF) units by $(2N - 8)$. Application specific integrated circuit (ASIC) and field programmable gate array (FPGA) results show that the proposed design for 1024-point achieves 31.54% less area, 30.13% less power, 33.56% less area-delay product (ADP), 27.11% less sliced look-up tables (SLUTs) and 28.37% less flip-flops (FFs) as compared to the best existing scheme.

Efficient Methods for Estimating Sinusoidal Frequencies Using Line Spectral Pairs

14:33 – 14:51

P. Vishnu (IIT Madras, India)

C.S. Ramalingam (IIT Madras, India)

The maximum likelihood (ML) method of estimating the frequencies of p sinusoids in the presence of AWGN is computationally very costly because of the dimensionality of the error surface; the advantage is that the ML method has the lowest threshold among all known practical estimators. We propose a low complexity method using Line Spectral Pairs (LSPs), where the LSPs are derived from an estimated $A(z)$ obtained using Multiple Signal Classification (MUSIC) method. The proposed method evaluates the likelihood function at significantly fewer number of points—at most $5pC_p$ —for getting the estimates. Furthermore, no iterative finer search is required. Nevertheless, the proposed method's threshold is comparable to that of ML when tested using the well-known two-sinusoids example; similar performance was observed in the case of three sinusoids. Further improvements were observed when the beamformer function was used for detecting and removing outliers. For the two-sinusoid case, outlier removal resulted in a threshold that was lower than that of ML by as much as 9 dB ($3\pi/2$ case). We also present results for a direction of arrival (DOA) estimation example that results in the same threshold as that of ML.

TECHNICAL PROGRAM

Adaptive Multiple-pixel Wide Seam Carving

14:51 –15:09

Diptiben Patel (IIT Gandhinagar, India)

Srivathsan Shanmuganathan (University of Jaffna, Sri Lanka & IIT Gandhinagar, India)

Shanmuganathan Raman (IIT Gandhinagar, India)

Content-aware image retargeting methods address the resizing of an image to be displayed on devices having different aspect ratios and resolutions. Seam carving method is an effective image retargeting method which suffers from high computational complexity. It requires one to find one-pixel wide minimum energy path in either vertical or horizontal direction, called seam, to reduce the image size by one pixel. In this paper, we propose an acceleration of the seam carving method by expanding the width of the seam making it multiple-pixel wide seam carving. The two types of energies: one corresponding to the pixels to be removed and another corresponding to the pixels across the multiple-pixel wide seam, increase as the width of the seam increases. In order to prevent the increase in these energies, we make the width of the seam adaptive as a function of the number of iterations. We find the width of a seam for each iteration as a prior for the seam carving process using a set of maximum energy seams in an orthogonal direction to the seam carving process. Qualitative and quantitative results prove that the proposed method performs faster and better than the other state-of-the-art image retargeting operators.

Design of Discrete Frequency-Coding Waveforms Using Phase-Coded Linear Chirp for Multiuser and MIMO Radar Systems

15:09 – 15:27

Arijit Roy (IIT Guwahati, India)

Debasish Deb (DRDO Bangalore, India)

Harshal B. Nemade (IIT Guwahati, India)

Ratnajit Bhattacharjee (IIT Guwahati, India)

Content-aware image retargeting methods address the resizing of an image to be displayed on devices having different aspect ratios and resolutions. Seam carving method is an effective image retargeting method which suffers from high computational complexity. It requires one to find one-pixel wide minimum energy path in either vertical or horizontal direction, called seam, to reduce the image size by one pixel. In this paper, we propose an acceleration of the seam carving method by expanding the width of the seam making it multiple-pixel wide seam carving. The two types of energies: one corresponding to the pixels to be removed and another corresponding to the pixels across the multiple-pixel wide seam, increase as the width of the seam increases. In order to prevent the increase in these energies, we make the width of the seam adaptive as a function of the number of iterations. We find the width of a seam for each iteration as a prior for the seam carving process using a set of maximum energy seams in an orthogonal direction to the seam carving process. Qualitative and quantitative results prove that the proposed method performs faster and better than the other state-of-the-art image retargeting operators.

TECHNICAL PROGRAM

Saliency Guided Image Detail Enhancement

15:27 – 15:45

Sanjay Ghosh (IISc Bangalore, India)

Ruturaj G. Gavaskar (IISc Bangalore, India)

Kunal N. Chaudhury (IISc Bangalore, India)

The use of visual saliency for perceptual enhancement of images has drawn significant attention. In this paper, we explore the idea of selectively enhancing salient regions of an image. Moreover, we develop an algorithm based on adaptive bilateral filtering for this purpose. In most of the filtering based methods, detail enhancement is performed by decomposing the image into base and detail layers; the detail layer is amplified and added back to the base layer to obtain the enhanced image. The decomposition is performed using edge-preserving smoothing such as bilateral filtering. The present novelty is that we use the saliency map to locally guide the smoothing (and the enhancement) action of the bilateral filter. The effectiveness of our proposal is demonstrated using visual results. In particular, our method does not suffer from gradient reversals and halo artifacts, and does not amplify fine details in non-salient regions that often appear as noise grains in the enhanced image. Moreover, if we choose to perform the filtering channelwise, then our method can be efficiently implemented using an existing fast algorithm

NW1: Communication Networks

Thursday, February 21, 14:15–15:45

Venue: MPC

Chair: Krishna P Jagannathan (IIT Madras, India)

Caching Partial Files for Content Delivery

14:15 – 14:33

Lakshmi Narayana V S CH (IIT Bombay, India)

Sambhav Jain (IIT Bombay, India)

Sharayu Moharir (IIT Bombay, India)

Numerous empirical studies have shown that users of video-on-demand platforms do not always watch videos in their entirety. A direct consequence of this is that not all parts of a video are equally popular. Motivated by this, we explore the benefits of dividing files into smaller segments for caching. We treat incoming requests as requests for segments of files and propose a Markovian request model which captures the time-correlation in requests. We characterize the fundamental limit on the performance of caching policies which only cache full files. Next, we propose and analyze the performance of policies which cache partial files. Using this, we characterize the potential for improvement in performance due to caching partial files and analyze its dependence on various system parameters like cache size and the popularity profile of the files being cached.

TECHNICAL PROGRAM

The HTTP/2 Server Push and Its Implications on Mobile Web Quality of Experience

14:33 – 14:51

Hema Kumar Yarnagula (IIT Guwahati, India)
Venkatesh Tamarapalli (IIT Guwahati, India)

In recent years, an unprecedented growth in the usage of mobile devices for web browsing poses a challenge for the service providers to assure the user-perceived quality. In the context of web quality of experience (QoE), quality perception is mostly dominated by the page load time (PLT). HTTP/2 protocol, with the server push feature, promises to address the design limitations of HTTP/1.1 that inhibit optimal web performance. However, it remains largely unclear if HTTP/2 can really improve web QoE for mobile browsing.

In this paper, we experimentally investigate the web QoE with HTTP/2. We assess the web QoE for several popular websites on a controlled testbed emulated with real 4G/LTE and 3G network traces. Our experiments investigate the impact of both network latency and packet loss ratio on the mobile web QoE. The results clearly show 24% improvement in the PLT, on an average, with HTTP/2 over mobile networks. However, we identify that HTTP/2 with server push is necessarily not the fail-safe solution for improving mobile web QoE under all conditions. We noticed that HTTP/2 loads the web pages slower than HTTP/1.1 when the network packet loss ratio is more than 2%. Our study could be used as the basis to derive a set of guidelines on the usage of the HTTP/2 server push to improve the end-user web QoE, especially in mobile devices.

A Graph Based Clustering and Preconditioning of V-MIMO Wireless Sensor Networks

14:51 – 15:09

Rakesh Mundlamuri (Shiv Nadar University, India)
Thangapandian B (Shiv Nadar University, India)
Vijay Kumar Chakka (Shiv Nadar University, India)
Srikanth Goli (Shiv Nadar University, India)

This paper presents a graph based methodology for increasing the channel capacity of Virtual-Multiple Input Multiple Output (V-MIMO) defined over a Wireless Sensor Network (WSN). A fully connected graph $G(V, E, W)$ is defined for a WSN. Then, we propose a new clustering algorithm based on the Fiedler vector of the graph G which divides the sensor nodes V into two clusters (transmitting and receiving antennas). The links between these two clusters results in V-MIMO network. Next, a Modified Maximum Spanning Tree Search algorithm (MMASTS) is proposed on V-MIMO to enhance the average channel capacity. Simulation performance of average channel capacity and uncoded Bit Error Rate (BER) are plotted using different precoding techniques like Zero Forcing (ZF) and Minimum Mean Square Error (MMSE). These are also used for comparing the performance of proposed Fiedler vector based clustering with k – means clustering.

TECHNICAL PROGRAM

Analysis of Mid-Haul Characteristics for LTE-NR Multi-Connectivity in Heterogeneous Cloud RAN

15:09 – 15:27

Ramakrishnan S (IIT Delhi, India)

Subrat Kar (IIT Delhi, India)

Dharmaraja Selvamuthu (IIT Delhi, India)

To address the capacity requirements resulting from huge growth in mobile data traffic, the mobile network operators (MNOs) are deploying heterogeneous Cloud based Radio Access Network (C-RAN) with a mix of Base stations viz. 4G Macro Base stations (offering better coverage) and 5G Small cells (offering better radio capacities). With LTE-NR Multi-Connectivity feature in 5G network, the User Equipments (UEs) can connect with both 4G and 5G simultaneously to take advantage of better coverage and capacity, so that its QoS is met appropriately. To maintain such simultaneous connections (with UE) across different BTS, the network must support stringent high bandwidth and lowlatency links (mid-haul links) between the Base station nodes. In a C-RAN environment, the inter Base station mid-haul links can be viewed as inter Virtual Machine (VM) communication link.

In this paper, we analyze the characteristics of the midhaul link for LTE-NR Multi-Connectivity feature (specifically E-UTRAN-NR Dual Connectivity EN-DC configuration) in a heterogeneous C-RAN deployment. We simulate the C-RAN scenario using NS3 network simulator with heterogeneous mix of 4G and 5G-NR Base stations with appropriate cloud architecture split. We use the BTS power model to derive the Computational Requirement (CR) of the Base station for vBBU (virtual Baseband Unit) placement algorithm in C-RAN. Through simulation, we review the bandwidth requirement of the mid-haul link and assess variation in the end-to-end delay when latency of mid-haul link is varied. From available cloud infrastructure hardware bench-mark results, we observe that the latency of inter VM communication links, varies depending on whether the two communicating VMs are on same or different Cloud servers. Thus, we propose "Neighbour Association-aware Placement" (NAP) algorithm for placement of the vBBUs in the same Cloud server and assess the benefits in the case of EN-DC configuration

ASER Analysis of General Order Rectangular QAM for Dual-Hop NLOS UV Communication System

15:27 – 15:45

Kamal K Garg (IIT Indore, India)

Praveen Kumar Singya (IIT Indore, India)

Vimal Bhatia (IIT Indore, India)

Ultraviolet (UV) communication is capable of providing non-line-of-sight (NLOS) wireless connectivity due to strong molecular and aerosol scattering experienced at UV wavelength. The effect of atmospheric turbulence in NLOS UV channel is usually ignored under the assumption of short distance communication. However, for long distance, relay assisted communication is commonly used. In this work, we consider a dual-hop amplify-andforward (AF) relayed outdoor NLOS UV communication system experiencing atmospheric turbulence, and derive closed-form expression of the average symbol error rate (ASER) for general order rectangular quadrature amplitude modulation (RQAM) scheme. The numerical values of ASER expression are compared with computer simulations to validate the accuracy of the theoretical analysis.

TECHNICAL PROGRAM

BP3: Best Paper Awards Session 3

Thursday, February 21, 16:15 – 17:45

Venue: GJH

Chair: Kalpana Dhaka (IIT Guwahati, India)

SSK Performance with SWIPT based Dual-Hop AF Relay over Rayleigh Fading

16:15- 16:33

Hemanta Kumar Sahu (IIT Bhubaneshwar, India)

P. R. Sahu (IIT Bhubaneshwar, India)

A cooperative communication system with space shift keying (SSK) modulation and simultaneous wireless information and power transfer (SWIPT) scheme is proposed. SWIPT can eliminate the need of external power supply at the relay whereas SSK modulation scheme reduces inter-channel interference, excludes inter antenna synchronization requirement and the number of radio frequency chains. An upper bound expression for the average bit error probability (ABEP) is obtained with multiple amplify-forward relays and a direct link from source node to destination node. Further, ABEP is analyzed for partial relay selection operation. Numerical and computer simulation results demonstrate performance improvement for SSK modulation combined with SWIPT.

Interference Violation Probability Constrained Underlay Cognitive Massive MIMO Network Under Imperfect Channel Knowledge

16:33- 16:51

Rama Gupta (IIT Guwahati, India)

Salil Kashyap (IIT Guwahati, India)

E. Venkata Pothan (IIT Guwahati, India)

We investigate the use of massive number of antennas at the cognitive base station (BS) in reducing interference caused to primary users (PUs) under imperfect channel knowledge without deteriorating the data rate provided to the cognitive user (CU). To this end, we develop a simple back-off factor based power adaptation policy for the cognitive BS which ensures that its transmissions do not violate the interference violation probability constraint at the PUs. We derive a new lower bound on the complement of the interference violation probability and also deduce a lower bound on the achievable rate of the CU when the cognitive BS has an imperfect estimate of its channels to the PUs and the CU. Through our analytical and numerical results, we quantify that the interference violation probability at the PUs can be reduced while providing a fixed data rate to the CU by deploying more number of cognitive BS antennas. Furthermore, if the number of PUs in the network increase, we show that the interference violation probability at the PUs and the data rate of the CU can be maintained at the same level by increasing the number of antennas at the cognitive BS.

TECHNICAL PROGRAM

Improved Tail Bounds for Missing Mass and Confidence Intervals for Good-Turing Estimator

16:51- 17:09

Prafulla Chandra (IIT Madras, India)

Aditya Pradeep (IIT Madras, India)

Andrew Thangaraj (IIT Madras, India)

The missing mass of a sequence is defined as the total probability of the elements that have not appeared or occurred in the sequence. The popular Good-Turing estimator for missing mass has been used extensively in language modeling and ecological studies. Exponential tail bounds have been known for missing mass, and improving them results in better confidence in estimation. In this work, we first show that missing mass is sub-Gamma on the right tail with the best-possible variance parameter under the Poisson and multinomial sampling models. This results in a right tail bound that beats the previously best known tail bound for deviation from mean up to about 0.2785. Further, we show that the sub-Gaussian approach cannot result in any improvement in the right tail bound for Poisson sampling. We derive confidence intervals for the Good-Turing estimator with better confidence levels and narrower width when compared to existing ones. Our results are worst case over all distributions.

Differential Phase Encoding Scheme for Measurement-Device-Independent Quantum Key Distribution

17:09 – 17:27

Shashank Kumar Ranu (IIT Madras, India)

Anil Prabhakar (IIT Madras, India)

Prabha Mandayam (IIT Madras, India)

This paper proposes a measurement-deviceindependent quantum key distribution (MDI-QKD) scheme based on differential phase encoding. The differential phase shift MDI-QKD (DPS-MDI-QKD) couples the advantages of DPS-QKD with that of MDI-QKD. The proposed scheme offers resistance against photon number splitting attack and phase fluctuations as well as immunity against detector side-channel vulnerabilities. The design proposed in this paper uses weak coherent pulses in a superposition of three orthogonal states, corresponding to one of three distinct paths in a delay-line interferometer. The classical bit information is encoded in the phase difference between pulses traversing successive paths. This 3-pulse superposition offers enhanced security compared to using a train of pulses by decreasing the learning rate of an eavesdropper and unmasking her presence with an increased error rate upon application of intercept and resend attack and beamsplitter attack. The proposed scheme employs phase locking of the sources of the two trusted parties so as to maintain the coherence between their optical signal, and uses a beamsplitter (BS) at the untrusted node (Charlie) to extract the key information from the phase encoded signals.

TECHNICAL PROGRAM

Truthful Double Auction Based VM Allocation for Revenue-Energy Trade-Off in Cloud Data Centers

17:27-17:45

Yashwant Singh Patel (IIT Patna, India)

Animesh Nighojkar (Medi-Caps Institute of Technology and Management, India)

Rajiv Misra (IIT Patna, India)

With the advances in virtualization technologies, cloud has emerged as a flexible and cost-effective service paradigm by provisioning on-demand VM resources to users via a pay-per-use business model. In cloud data centers, effective resource provisioning is required with the aim of minimizing energy consumption and maximizing cloud provider's revenue. However, the existing mechanisms have either focused on the optimization of energy, or the profit of cloud service provider (CSP) while incurring inefficient resource allocation. Thus to address these fundamental research challenges and to balance the trade-off between energy and revenue, we propose a Vickrey-Clarke-Groves (VCG) based truthful double auction mechanism (TDAM). In this paper, first, we have formulated a joint optimization problem and prove it NP-hard by reducing it to a multi-dimensional bin-packing problem. Then we design TDAM, a truthful double auction scheme and propose an efficient winning bid algorithm for VM allocation and a VCG based mechanism for calculating payment of each bid. Being a double auction, TDAM allows both the buyers (VMs) and the sellers (PMs) to submit their bids and asks respectively, and performs allocation based on the energy consumption, while upholding truthfulness, in order to avoid falsification of the submitted bid or ask values. Through theoretical analysis and extensive experiments we show that the TDAM makes a significant contribution while maintaining truthfulness, individual rationality, economic efficiency, and has polynomial time complexity.

ML1: Applications of Machine Learning

Thursday, February 21, 16:15 - 17:45

Venue: MPC

Chair: Ramakrishnan Ganesan (IISc Bangalore, India)

Development of Assamese Text-to-Speech System Using Deep Neural Network

16:15 – 16:33

Abhash Deka (IIT Guwahati, India)

Priyankoo Sarmah (IIT Guwahati, India)

Samudravijaya K (IIT Guwahati, India)

Mahadeva Prasanna (IIT Dharwad, India)

This paper describes the development of a text-to-speech system for Assamese language, using Deep Neural Network (DNN). The system is trained with speech data, collected by a consortium, that is available free of cost for academic use. The DNN based method eliminates the need for a grapheme to phoneme conversion; rather, it synthesizes speech directly from the UTF-8 based Assamese script. The results of objective and subjective evaluations confirm that the Assamese speech synthesized using DNN approach is better than the ones synthesized using the traditional hidden Markov model based text-to-speech system.

TECHNICAL PROGRAM

Multimodal Fusion of Speech and Text Using Semi-supervised LDA for Indexing Lecture Videos

16:33 – 16:51

Moula Husain (VTU, India)

Meena S M (B. V. Bhoomaraddi College of Engineering and Technology, India)

Lecture videos are the most popular learning materials due to their pedagogical benefits. However, accessing a topic or subtopic of interest requires manual examination of each frame of the video and it is more tedious when the volume and length of videos increases. The main problem thus becomes the efficient automatic segmentation and indexing of lecture videos that enables faster retrieval of specific and relevant content. In this paper, we present automatic indexing of lecture videos using topic hierarchies extracted from slide text and audio transcripts. Indexing videos based on slide text information is more accurate due to higher character recognition rates but, text content is very abstract and subjective. In contrast to slide text, audio transcripts provide comprehensive details about the topics, however retrieval results are imprecise due to higher WER. In order to address this problem, we propose a novel idea of fusing complementary strengths of slide text and audio transcript information using semi-supervised LDA algorithm. Further, we strive to improve learning of the model by utilizing words recognized from video slides as seed words and train the model to learn the distribution of video transcriptions around these seed words. We test the performance of proposed multimodal indexing scheme on 500 number of class room videos downloaded from Coursera, NPTEL and KLETU (KLE Technological University) classroom videos. The proposed multimodal fusion based scheme achieves an average percentage improvement of 44.49% F-Score compared with indexing using unimodal approaches.

Emotion Recognition from Varying Length Patterns of Speech Using CNN-based Segment-Level Pyramid Match Kernel Based SVMs

16:51 – 17:09

Shikha Gupta (IIT Mandi, India)

Kishalaya De (MIT Manipal, India)

Dileep Aroor Dinesh (IIT Mandi, India)

Veena Thenkanidiyoor (NIT Goa, India)

Convolutional Neural Networks (CNNs) and its variants have achieved impressive performance when used for different speech processing tasks like spoken language identification, speaker verification, speech emotion recognition, etc. Conventionally, CNNs for speech applications consider input features from fixed duration speech segments as input. In this work, we attempt to consider features from complete speech signal as input to CNN. We propose to use spatial pyramid pooling (SPP) layer in CNN architecture to remove the fixed length constraint and to consider features from varying length speech signals as input to CNN for an end to end training. Proposed architecture also results in varying size set of feature maps from convolution layer. Further, we propose novel CNNbased segment-level pyramid match kernel (CNN-SLPMK) as dynamic kernel between a pair of varying size set of feature maps for the classification framework using support vector machines (SVMs) based classifier. We demonstrate that our proposed approach achieves comparable results with state-of-the-art techniques for speech emotion recognition task.

TECHNICAL PROGRAM

Full-Reference Video Quality Assessment Using Deep 3D Convolutional Neural Networks

17:09 – 17:27

Sathya Veera Reddy Dendi (IIT Hyderabad, India)

Gokul Krishnappa (IIT Hyderabad, India)

Sumohana Channappayya (IIT Hyderabad, India)

We present a novel framework called Deep Video QQuality Evaluator (DeepVQUE) for full-reference video quality assessment (FRVQA) using deep 3D convolutional neural networks (3D ConvNets). DeepVQUE is a complementary framework to traditional handcrafted feature based methods in that it uses deep 3D ConvNet models for feature extraction. 3D ConvNets are capable of extracting spatio-temporal features of the video which are vital for video quality assessment (VQA). Most of the existing FRVQA approaches operate on spatial and temporal domains independently followed by pooling, and often ignore the crucial spatio-temporal relationship of intensities in natural videos. In this work, we pay special attention to the contribution of spatio-temporal dependencies in natural videos to quality assessment. Specifically, the proposed approach estimates the spatio-temporal quality of a video with respect to its pristine version by applying commonly used distance measures such as the L_1 or the L_2 norm to the volume-wise pristine and distorted 3D ConvNet features. Spatial quality is estimated using off-the-shelf full-reference image quality assessment (FRIQA) methods. Overall video quality is estimated using support vector regression (SVR) applied to the spatio-temporal and spatial quality estimates. Additionally, we illustrate the ability of the proposed approach to localize distortions in space and time.

Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling

17:27 – 17:45

Sambit Behura (NIT Rourkela, India)

Shrishail M Hiremath (NIT Rourkela, India)

Sarat Kumar Patra (NIT Rourkela, India)

Siddharth Deshmukh (NIT Rourkela, India)

Subham Kedia (NIT Rourkela, India)

Deep learning techniques have recently exhibited unprecedented success in classification problems with ill-defined mathematical models. In this paper, we apply deep learning for RF data analysis and classification. We present a novel method of using I-Q time samples to form images with 'Time and Discrete Orthonormal Stockwell Transform Domain Channels' which are used for training a convolutional neural network (CNN) for radio modulation classification. Also, a concept inspired from transfer learning is used in extending the number of output classes of the CNN, which helps the network to estimate the approximate SNR of the input signal as well and further improve the classification accuracy. Such a network trained on Time and Stockwell Channeled Images performs superior to similar networks that are trained on just raw I-Q time series samples or timefrequency images, especially when training samples are less. The network achieved an overall classification accuracy of 97.3% at 8 dB SNR over a class of 10 radio modulation schemes (for both digital and analog systems). The study shows that such a trained network can be well applied to achieve high classification accuracies at low and moderate SNR scenarios.

TECHNICAL PROGRAM

COM1: Communications Algorithms and Implementations

Friday, February 22, 11:15 - 12:45

Venue: GJH

Chair: Amit Dutta (IIT Kharagpur, India))

Sparse Bayesian Learning (SBL)-Based Frequency-Selective Channel Estimation for Millimeter Wave Hybrid MIMO Systems

11:15- 11:33

Suraj Srivastava (IIT Kanpur)

Suraj Kumar Patro (IIT Kanpur)

Aditya K. Jagannatham (IIT Kanpur)

Govind Sharma (IIT Kanpur)

This work develops a novel sparse Bayesian learning (SBL)-based channel estimation technique for frequency-selective millimeter wave (mmWave) multiple-input multiple-output (MIMO) systems. Towards this end, the concatenated frequency-selective MIMO channel matrix is represented in terms of the beamspace channel vector employing suitable transmit and receive array response dictionary matrices. Subsequently, a multiple measurement vector (MMV) model is developed for estimation of the sparse beamspace channel vector considering the block transmission of zero-padded training frames. The unique aspects of the proposed scheme are that it exploits the groupsparsity inherent in the equivalent beamspace channel vector of the frequency-selective mmWave MIMO channel and also considers the effect of correlated noise in the equivalent system model due to RF-combining. This feature, coupled with the improved ability of SBL for sparse signal recovery, leads to a significantly enhanced performance of the proposed scheme in comparison to the orthogonal matching pursuit (OMP) technique proposed recently. Bayesian Cramér-Rao bounds (BCRBs) are also derived to characterize the estimation performance. Simulation results are presented to demonstrate the improved performance of the proposed SBL-based channel estimation technique in comparison to the existing scheme and also a performance close to the various benchmarks.

MIMO-FBMC Channel Estimation with Limited, and Imperfect Knowledge of Channel Correlations

11:33- 11:51

Prem Singh (IIT Kanpur, India)

K. Vasudevan (IIT Kanpur, India)

This paper presents and analyses the performance of training-based least squares (LS) and minimum mean square error (MMSE) channel estimation schemes for multiple input multiple output (MIMO) filter bank multicarrier (FBMC) systems based on the offset quadrature amplitude modulation (OQAM) in the presence of limited, and imperfect knowledge of the channel correlations. First, a linear MMSE (LMMSE) technique for MIMO-FBMC channel estimation, which require a priori knowledge of channel correlation matrix, is examined by utilizing the second-order statistical properties of the intrinsic interference in FBMC systems. A biased LS (BLS) and relaxed LMMSE (RLMMSE) MIMO-FBMC channel estimation schemes, which require prior knowledge of the trace of the channel correlation matrix, are proposed. The LS-BLS and LS-RLMMSE schemes for MIMO-FBMC channel estimation are investigated in the presence of imperfect knowledge of the channel correlations. The mean square error is derived for the proposed schemes by exploiting statistical properties of the intrinsic interference. Simulation results show that the proposed schemes present an excellent trade-off between the achieved performance and required a priori knowledge of the channel correlations.

TECHNICAL PROGRAM

Hardware Implementation of Filtered OFDM for BB-PLC using Software Defined Radio

11:51- 12:09

Sumesh K P (NIT Goa, India)

Ankit Dubey (NIT Goa, India)

Trilochan Panigrahi (NIT Goa, India)

Among several in-home communication systems, broadband over power line communication (BB-PLC) provides better connectivity at cheaper installation cost without disturbing the existing infrastructure. Conventional broadband systems, including BB-PLC, use the orthogonal frequency division modulation (OFDM) to combat the issue of frequency selective fading at the cost of extra bandwidth. However, it is proposed and shown that the spectral efficiency of the OFDM can be improved either by using filter bank multi-carrier (FBMC) or filtered OFDM (FOFDM). In the former, all the sub-carriers of an OFDM symbol are filtered, however, in the latter, only the final OFDM symbol is filtered. Thus, it is a cost-effective solution to use the F-OFDM over the FBMC. This paper presents the hardware implementation of the F-OFDM for BB-PLC. Different prototype low-pass filters are used for comparative analysis, especially for the BB-PLC. The simulations are carried out in National Instruments' LabVIEW software and the hardware implementation of the transceiver is done on National Instruments' Software Defined Radio (SDR) set called USRP 2920. Polycab 1.5 SQ mm home wiring power cable is used as the communication channel in testing the BB-PLC system. From the power spectral density analysis, it is concluded that the F-OFDM has better spectral efficiency than the conventional OFDM. Further, it is observed that as the length of the filter increases the better spectral efficiency is achieved but at the cost of increased complexity.

Codebook based Precoding for Multiuser MIMO Broadcast Systems: An MM Approach

12:09 – 12:27

Sai Subramanyam Thoota (IISc Bangalore, India)

Prabhu Babu (IIT Delhi, India)

Chandra R. Murthy (IISc Bangalore, India)

The goal of this paper is to propose a novel, principled approach to solve non-convex optimization problems that arise in multiuser (MU) multiple input multiple output (MIMO) cellular wireless communication systems. We explore a minorization-maximization (MM) optimization approach, which is guaranteed to converge to a stationary point starting from any initialization. One of the important problems in wireless communications is sum rate maximization in MU MIMO broadcast systems, in which multiple data streams are simultaneously transmitted to all users. In this paper, we adopt a codebook based precoding method, where each data stream is beamformed using a vector selected from a predetermined codebook. Our objective is to determine the selection of beamforming vectors and power allocation to each beam to maximize the achievable sum rate. We reformulate the problem to facilitate the application of MM procedure in a nested fashion. The outcome is a novel, iterative, and computationally efficient solution, which we call the inverseMM (IMM) algorithm. We illustrate the superior performance of our algorithm compared to existing approaches through Monte Carlo simulations. The advantages of computational efficiency, simple implementation, and structured approach makes the MM framework a good candidate for solving non convex optimization problems in wireless communications.

TECHNICAL PROGRAM

Modified Generalised Quadrature Spatial Modulation

12:27 – 12:45

Kiran Gunde (IISc Bangalore, India)

K.V.S. Hari (IISc Bangalore, India)

In this paper, we propose a Modified Generalised Quadrature Spatial Modulation (mGQSM) scheme with multiple RF chains. The proposed scheme, compared to GQSM, proposes a novel codebook design which provides one extra bit per channel use (bpcu) spectral efficiency with the constraint of $\{\log_2 N_t N_{rf}\} \geq 0.5$, where N_t denotes number of transmit antennas, and N_{rf} denotes number of RF chains, $1 \leq N_{rf} \leq b N_t / 2$. Using the ML detection algorithm, we study the performance of mGQSM with and without imperfect channel state information, via numerical simulations. We compute the computational complexity of ML-decoding in terms of real valued multiplications and introduce a variant of mGQSM called Reduced Codebook mGQSM (RC-mGQSM) to reduce the complexity but resulting in a decrease in spectral efficiency.

RF2: RF & Microwaves

Friday, February 22, 11:15 - 12:45

Venue: MPC

Chair: Amit Deshmukh (DJSCOE, India)

Predistortion Linearizer Design for K_u Band RF Power Amplifier

11:15 - 11:33

Girish Chandra Tripathi (IIT Roorkee, India)

Meenakshi Rawat (IIT Roorkee, India)

This paper presents a K_u band analog predistorter linearizer for improving the linearity of high-power radio frequency (RF) amplifiers. This method is applicable for both solid-state power amplifiers (SSPA) and traveling-wave tube amplifiers (TWTA). The designed linearizer consists of analog components, hence it is wideband as compared to the digital predistortion. Moreover, due to most of the passive components, the circuit is simpler and with the advantage of individual control of amplitude modulation to amplitude modulation (AM-AM) and amplitude modulation to phase modulation (AM-PM) conversions. For proof of concept, the designed linearizer is simulated with ZX60-14012L+ class AB SSPA. The S2D SSPA model is extracted using vector network analyzer. The proposed linearizer shows a reduction in third order intermodulation of 37 dB approximately at 4 dB output power backoff for two-tone signal. Similarly, for Long-Term Evolution (LTE) 20 MHz signal after linearization adjacent channel power ratio is approximately 47 dBc and shows a correction of 16 dB.

TECHNICAL PROGRAM

LSTM-Deep Neural Networks Based Predistortion Linearizer for High Power Amplifiers

11:33 - 11:51

Meenakshi Rawat (IIT Roorkee, India)

Deepmala Phartiyal (IIT Roorkee, India)

Linear high power amplifiers (HPAs) are the need of current communications technology. But, almost all PAs show non-linear characteristics during amplification which are reflected in the transmitted signal in the form of distortions. Linearization is a process to suppress the effect of the non-linear characteristic of a PA. Various methods are available to perform linearization. Predistortion (PD) linearization methods are very successful due to its simplicity in design and ease of integration with PAs. PD linearization methods observe the PA dynamic characteristics (nonlinearity) and then formulate an "inverse transfer function" to suppress this non-linearity. In the last decade, machine learning (ML) based PD linearizers are proposed and proved useful. Since then, numerous ML-PD linearizers have been developed. Shallow neural networks (NNs) based PD linearizers are successfully used to map the inverse transfer function but lack generalization performance in the presence of system conditions (IQ imbalance, DC offset). With deep learning (DL) technology, deep neural networks (DNNs) can map the complex inverse transfer function under different system conditions. This study proposes a long short-term memory (LSTM) DNN based PD linearizer for linearization of PA under different conditions. In this study, it is shown that LSTM is able to extract and exploit memory effects of PA over the perceptron. Comparative results with shallow NNs suggest reliable potential in the proposed DNN model in terms of generalization performance.

Design of Multiband Negative Permittivity Metamaterial Based on Interdigitated and Meander Line Resonator

11:51 – 12:09

Rohan Deshmukh (VNIT, Nagpur, India)

Dushyant Marathe (VNIT, Nagpur, India)

Kishore Kulat (VNIT, Nagpur, India)

We report a new design of multiband electric metamaterial resonator based on integration of interdigitated structure and meander line with square ring. This metamaterial resonator has three distinct electric resonances and negative permittivity regions at C, X band of frequencies. The scattering parameters of proposed sub-wavelength resonator are analysed using full wave electromagnetic simulator Ansys HFSS to demonstrate the presence of electric response at resonant frequencies within 2-12 GHz band. Effective medium parameters permittivity, permeability and refractive index are extracted from simulated scattering parameters. The investigations are also carried out regarding independence of magnetic dipolar activity on flow of surface current. Performance comparison of proposed resonator with single negative SNG ($\epsilon < 0$ and/or $\mu < 0$) resonators is carried out.

TECHNICAL PROGRAM

Design of Frequency-Signature Based Multiresonators Using Quarter Wavelength Open Ended Stub for Chipless RFID Tag

12:09 - 12:27

P Prabavathi (PSG College of Technology, India)

Subha Rani S (PSG College of Technology, India)

A quarter wavelength open stub multiresonators are proposed for chipless Radio Frequency Identification (RFID) tag. The data capacity of the tag is 10 bit and open stub resonator operates in the frequency band of 2 GHz to 4 GHz. The data in the tag is encoded using absence or presence coding and frequency shift coding (FSC). The data stored is read and transmitted using the planar circular patch monopole ultra-wide band (UWB) antenna. The tag consists of two cross polarized sending and receiving planar circular patch (PCP) monopole UWB antennas connected to the multiresonators. The span of the multiresonator based chipless RFID tag is 23.8mmx17mm which is designed on FR4 substrate with dielectric permittivity of 4.4 and tangent loss of 0.01. It is designed and tested under simulation using ADS software and the vector network analyzer (VNA) after fabrication. The tag has the insertion loss in the range of -10 dB to -30 dB and a bit density of 2.47bits/cm².

Slot Antenna Miniaturization Using Copper Coated Circular Dielectric Material

12:27 - 12:45

Khan Masood Parvez (Aliah University, India)

Enamul Khan (Aliah University, India),

SK. Moinul Haque (Aliah University, India)

Jinia Aktar (Aliah University, India)

The contribution of this paper is to propose a simple slot antenna miniaturization method using copper coated FR_4 dielectric material loading technique. The operating frequency for reference antenna and loaded antenna are 2.86GHz and 1.66GHz respectively. As a result, overall size of proposed antenna reduces by a ratio of 41.95%. A parametric study on various copper coated dielectric materials is presented to better understand the effect of the permittivity on slot antenna miniaturization. The antenna topologies are designed and analyzed using High Frequency Structure simulator (HFSS) tool. The prototype was fabricated and measured, and the measured results show good agreements with the simulated ones. This antenna can be very useful for various wireless communication systems.

TECHNICAL PROGRAM

COM2: Optical and Quantum Communications

Friday, February 22, 13:45 - 15:15

Venue: GJH

Chair: Varun Raghunathan (IISc Bangalore, India)

Quantum Random Number Generator with One and Two Entropy Sources

13:45 - 14:03

Gautam Shaw (IIT Madras, India)

Sivaram SR (IIT Madras, India)

Anil Prabhakar (IIT Madras, India)

Quantum random number generators (QRNGs) are an integral part of quantum key distribution (QKD) systems. To better understand the inherent physical processes, we compare the random numbers generated by two separate schemes, one is based on entropy (arrival time of photons) and another with an additional source of entropy (space) i.e, path superposition of arrival time of photons from a weak coherent source on a gated InGaAs single photon detector.

Both experiments yield bits that appear random. However, they satisfy different criteria of randomness. The weak coherent source has a Poissonian distribution and extracting the variation about the arrival time of photons on gated SPD yields a source of random numbers that pass most of the Dieharder Tests. With the inclusion of superposition, we obtain random numbers that pass all the Dieharder tests. The physical origins of the random numbers in the two experiments is different, one is single entropy source based and other one is two entropy source based, and this is reflected in the outcomes of the different tests for randomness.

Simulation of Emission Wavelength of Quantum Dot Based Single Photon Sources

14:03 - 14:21

Jyothish M (IIT Madras, India)

Fredy Francis (IIT Madras, India)

Manivasakan R (IIT Madras, India)

Advances in quantum information processing and the requirements of quantum key distribution schemes have made high quality single photon sources, extremely essential. Here generation of a single photon from a semiconductor quantum dot using a semi-classical approach is investigated.

Finite Element Method is used for the Eigen mode analysis of a typical pyramidal semiconductor quantum dot [3]. A design methodology is also proposed for obtaining the required emission wavelengths. Additionally, effects of wetting layer, height to base ratio and strain due to lattice mismatch are investigated. An Empirical relationship is obtained between pyramid geometry and emission wavelength. The simulation results were verified against the experimental works including [7] in $1.2 \mu\text{m}$ to $1.3 \mu\text{m}$ emission regime, and good match was observed.

TECHNICAL PROGRAM

Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier

14:21 – 14:39

Mehul Anghan (IIT Mumbai, India)

Rashmi Kamran (IIT Mumbai, India)

Nihar Gulati (IIT Kanpur, India)

Nandakumar Nambath (IIT Goa, India)

Shalabh Gupta (IIT Bombay, India)

Self-homodyne systems with polarization multiplexed carrier offer an local oscillator-less (LO-less) coherent receiver with simplified signal processing requirement that can be a good candidate for high-speed short-reach data center interconnects. The practical implementation of these systems is limited by the requirement of polarization control at the receiver end for separating the carrier and the modulated signal. In this paper, effect of polarization impairments in polarization diversity based systems is studied and modeled. A novel and practical adaptive polarization control technique based on optical power feedback from one polarization is proposed for polarization multiplexed carrier based systems and verified through simulation results. The application of the proposed concept is also experimentally demonstrated for a quadrature phase shift keying (QPSK) system with polarization multiplexed carrier.

Analysis of Beam Wander Effect of Flat-topped Multi-Gaussian Beam for FSO Communication Link

14:39 - 14:57

Arka Mukherjee (IIT Delhi, India)

Subrat Kar (IIT Delhi, India)

V K Jain (IIT Delhi, India)

Atmospheric turbulence causes severe impairment of FSO communication link. By using the earlier model for the Gaussian beam, we analyze the beam wander effect for the flat-topped multi-Gaussian beam. The link availability decreases drastically in high turbulence regime for a Gaussian beam due to turbulence induced beam wander. In this paper, we model each turbulent eddy as a thin dielectric lens with Gaussian shaped refractive index profile and assume there are several sheets of eddies throughout the propagation path. We consider uniformly distributed eddy positions in a laminar sheet with Chi-Square distributed eddy sizes. We graphically demonstrate beam wander characteristics for different beam sizes and orders of the flat-topped multi-Gaussian beam in all three turbulence regimes characterized by different refractive index structure parameter values. Our results show that the flat-topped beam has a limited advantage in weak and moderate turbulence regimes. But it has a significant advantage in high turbulence regime to mitigate link outage due to beam wander.

TECHNICAL PROGRAM

Qubits through Queues: The Capacity of Channels with Waiting Time Dependent Errors

14:57 - 15:15

Krishna P Jagannathan (IIT Madras, India)

Avhishek Chatterjee (IIT Madras, India)

Prabha Mandayam (IIT Madras, India)

We consider a setting where qubits are processed sequentially, and derive fundamental limits on the rate at which classical information can be transmitted using quantum states that decohere in time. Specifically, we model the sequential processing of qubits using a single server queue, and derive explicit expressions for the capacity of such a 'queue-channel.' We also demonstrate a sweet-spot phenomenon with respect to the arrival rate to the queue, i.e., we show that there exists a value of the arrival rate of the qubits at which the rate of information transmission (in bits/sec) through the queue-channel is maximised. Next, we consider a setting where the average rate of processing qubits is fixed, and show that the capacity of the queue-channel is maximised when the processing time is deterministic. We also discuss design implications of these results on quantum information processing systems.

SP2: Speech Segmentation and Detection

Friday, February 22, 13:45 - 15:15

Venue: Room MPA

Chair: Thippur Sreenivas (IISc Bangalore, India)

Detection of Vowel-Like Speech Using Variance of Sample Magnitudes

13:45 - 14:03

Nagapuri Srinivas (NIT Patna, India)

Gayadhar Pradhan (NIT Patna, India)

Kishore Kumar Puli (NIT Andhra Pradesh, India)

Vowel, semi vowel and diphthong sound units are collectively referred to as vowel-like speech (VLS). VLS are dominant voiced regions in a given speech signal. Consequently, within a short-analysis frame the variance of sample magnitudes (VSM) is significantly higher for VLS when compared with other speech regions. In this work, a signal processing approach is proposed to robustly extract the VSM within an analysis frame. The VSM at each time instant is then non-linearly mapped (NLM) using negative exponential function to suppress the fluctuations. The NLM-VSM values are nearly constant and significantly less in magnitude for VLS than other speech, silence and noise regions. The NLM-VSM is used as a front-end feature for detecting the VLS in a given speech signal. The experimental results presented in this paper show that, for clean as well as noisy speech signals, the proposed feature outperforms some of the earlier reported features for the task of detecting VLS and corresponding onset and offset points.

TECHNICAL PROGRAM

Detection of Vowels in Speech Signals Degraded by Speech-Like Noise

14:03 - 14:21

Avinash Kumar (NIT Patna, India)

Sarmila Garnaik (Veer Surendra Sai University of Technology, India)

Ishwar Chandra Yadav (NIT Patna, India)

Gayadhar Pradhan (NIT Patna, India)

Syed Shahnawazuddin (NIT Patna, India)

Detecting vowels in a noisy speech signal is a very challenging task. The problem is further aggravated when the noise exhibits speech-like characteristics, e.g., babble noise. In this work, a novel front-end feature extraction technique exploiting variational mode decomposition (VMD) is proposed to improve the detection of vowels in speech data degraded by speech-like noise. Each short-time analysis frame of speech is first decomposed into a set of variational mode functions (VMFs) using VMD. The logarithmic energy present in each of the VMFs is then used as the front-end features for detecting vowels. A three-class classifier (vowel, non-vowel and silence) with acoustic modeling based on long short-term memory (LSTM) architecture is developed on the TIMIT database using the proposed features as well as mel-frequency cepstral coefficients (MFCC). Using the three-class classifier, frame-level time-alignments for a given speech utterance are obtained to detect the vowel regions. The proposed features result in significantly improved performance under noisy test conditions than the MFCC features. Further, the vowel regions detected using the proposed features are also quite different from those obtained through the MFCC. Exploiting the aforementioned differences, the evidences are combined to further improve the detection accuracy.

Modelling Glottal Flow Derivative Signal for Detection of Replay Speech Samples

14:21 - 14:39

Jagabandhu Mishra (IIT Dharwad, India)

Debadatta Pati (NIT Nagaland, India)

Mahadeva Prasanna (IIT Dharwad, India)

It is a widely known fact that automatic speaker verification systems are quite vulnerable to replay speech. The present work deals with detecting replay speech by using the information available in glottal flow derivative (GFD) signal. In signal processing terms, the speech signal can be represented as the response of a vocal-tract system with excited by a excitation source in the form of glottal flow. The effect of record and replay devices distorted the spectral characteristics of the naturally uttered speech sample, resulting distortion in corresponding GFD signals. In this work the GFD signals are parameterized by using standard mel filters and Gaussian mixtures models are made for detection.

Although various methods are available, by correlation analysis it is observed that in the context of the present work the dynamic programming phase slope algorithm (DYPSA) method is relatively more effective in estimating the GFD signals. The experimental studies are made on ASVSpoof2017 database. The proposed glottal flow derivative mel frequency cepstral coefficients (GFDMFCC) feature provides 20.53% equal error rate (EER). This performance is comparatively poor than by speech and residual based features. It is mainly due to the absence of fine structure information in estimated GFD signal. However, in fusion with speech signal based constant-Q cepstral coefficients (CQCC) features, the GFDMFCC feature provides an improvement of 10.30% with reference to conventional residual feature. This shows the usefulness of modelling GFD signals for detection of replay signals.

TECHNICAL PROGRAM

Comparison of Low Dimension Speech Segment embeddings: Application to Speaker diarization

14:39 - 14:57

Srikanth Raj Chetupalli (IISc Bangalore, India)

Thippur Sreenivas (IISc Bangalore, India)

Anand Gopalakrishnan (NIT Surathkal, India)

Segment clustering is a crucial step in unsupervised speaker diarization. Bottom-up approaches, such as, hierarchical agglomerative clustering technique are used traditionally for segment clustering. In this paper, we consider the top-down approach to clustering, in which a speaker sensitive, low-dimensional representation of segments (speaker space) is obtained first, followed by Gaussian mixture model (GMM) based clustering. We explore three methods of obtaining the low dimension segment representation: (i) multi-dimensional scaling (MDS) based on segment to segment stochastic distances (ii) traditional principal component analysis (PCA), and (iii) factor analysis (i-vectors), of GMM mean super-vectors. We found that, MDS based embeddings result in better representation and hence result in better diarization performance compared to PCA and even i-vector embeddings.

Improved Epoch Extraction from Speech Signals Using Wavelet Synchrosqueezed Transform

14:57 - 15:15

Govind D (Amrita Vishwa Vidyapeetham, India)

S Lakshmi Priya (Amrita Vishwa Vidyapeetham, India)

Akarsh S (Amrita Vishwa Vidyapeetham, India)

Ganga Gowri B (Amrita Viswa Vidyapeetham, India)

Soman K P (Amrita Vishwa Vidyapeetham, India)

Synchrosqueezed wavelet transform (WSST) is an effective tool in tracking instantaneous frequency of a given signal. The objective of the present work is to propose a WSST based method for accurate epoch estimation from speech. Epochs in speech represent the instants where the excitation to the vocal-tract is maximum and instantaneous F0 contour is derived from epoch locations. The proposed hypothesis in this paper is that the signal reconstructed by discarding higher frequency modes (above the mean F0) in the WSST transformed time frequency domain observed to predominantly represent source characteristics. The presence of the source characteristics in the modified WSST reconstructed signal is validated by the improved identification accuracy obtained for the epochs estimated from clean speech utterances of CMU-Arctic database. To further demonstrate the effectiveness of the WSST in improving the overall epoch estimation performance, a WSST modified zero frequency filtering (ZFF) of speech, which is one of the simple and effective tools for epoch extraction, is proposed. The sharp instantaneous frequency representation by WSST also found to be effective in estimating epochs emotion utterances where rapid pitch variations are present. The improved epoch estimation performance from emotive utterances are confirmed by validating on the German emotion speech corpus (EmoDb).

TECHNICAL PROGRAM

COM3: Communication Theory

Friday, February 22, 15:45-17:15

Venue: GJH

Chair: Salil Kashyap (IIT Guwahati, India)

Performance Analysis and Optimization of Interference Limited Multi-Antenna BRN

15:45 - 16:03

Imtiyaz Khan (NIT Rourkela, India)

Dhulipudi Krishna Kanth (NIT Rourkela, India)

Poonam Singh (NIT Rourkela, India)

This paper investigates the outage performance of multiple antenna bidirectional relaying network (BRN) in the presence of co-channel interference (CCI). Herein, multi-antenna sources exchange information bi-directionally with the help of a single-antenna relay terminal. Under such scenario, we evaluate and compare the performance of two amplify-and-forward based multi-antenna transmission strategies viz., beamforming (BF) and antenna selection (AS). We derive the tight upper bound expressions of end outage probability (OP) for both the strategies over Rayleigh fading channel. We further conduct asymptotic analysis to examine the achievable diversity order of the considered system. To gain more insights, we analyze the power optimization problem to minimize the OP for different scenarios. Finally, Monte-Carlo simulation results are given to attest our theoretical analysis. Our finding suggests that the BF overperform the AS scheme at the expense of additional complexity.

Outage Analysis of an Asymmetric Dual Hop PLC-VLC System for Indoor Broadcasting

16:03 - 16:21

Manan Jani (Netaji Subhas Institute of Technology, India)

Parul Garg (Netaji Subhas Institute of Technology, India)

Akash Gupta (Netaji Subhas Institute of Technology, India)

We propose and investigate the performance of a novel asymmetric dual hop relay based power line communication (PLC) and visible light communication (VLC) system for the purpose of indoor broadcasting. The PLC link experiences log-normal fading and additive noises. The PLC link is used as a back-haul link for the VLC downlink transmission system which serve multiple end users. The characteristics of the VLC links are dependent on the random position of the end users. Novel closed form expressions for the cumulative distribution function (CDF) of the end to end signal to noise ratio (SNR) of the proposed system is derived. Also, capitalising on these derived entities, the system's performance is evaluated in terms of the outage probability (OP) metric.

TECHNICAL PROGRAM

Error Performance of QAM GFDM Waveform with CFO under AWGN and TWDP Fading Channel

16:21 - 16:39

Sapta Girish Babu Neelam (Bharat Electronics Limited & IIT Bhubaneswar, India)

Pravas Ranjan Sahu (IIT Bhubaneswar, India)

Generalized frequency division multiplexing (GFDM) is a strong contender waveform for 5G cellular communications. This letter derives closed form symbol error rate (SER) expressions with carrier frequency offset (CFO) in the presence of additive white gaussian noise (AWGN) channel for 1. Normal BPSK, QPSK and 16-QAM-GFDM and 2. Time shift and Frequency shift, offset QPSK (OQPSK) and offset 16-QAM GFDM waveforms. This letter also derives SER expressions under TWDP fading channel for BPSK and QPSK GFDM waveforms. It is observed that under AWGN channel 1. SER performance of frequency shift offset QAM (FS-OQAM) GFDM is better than conventional MF, ZF and MMSE receivers and 2. FS-OQAM-GFDM slightly outperforms time shift OQAM (TS-OQAM) GFDM at higher values of CFO. The performance analysis studies show that TWDP fading can result in a performance poorer than Rayleigh fading when two specular components which are equal in strength and antiphase cancel each other for $K > 6$ dB. The numerically evaluated results are in close agreement with the simulated results.

Performance Analysis of Wireless Powered Decode-and-Forward Relay System

16:39 - 16:57

Pawan Kumar (IIT Guwahati, India)

Kalpana Dhaka (IIT Guwahati, India)

We consider a two-hop decode-and-forward relay system with wireless powered source node. Source and destination are in the coverage range of the relay and direct link connecting them is assumed to be blocked. The signal transmitted by the relay node is used to harvest energy at the source and forward data to the destination in even time slot. In the following odd slot the energy harvested at source is used to communicate data to relay node. The cycle continues and the source data is communicated to the blocked destination node. The average symbol error rate (SER) of the system is analyzed for Rayleigh faded links when data is i) M-ary phase-shift keying modulated with coherent detection and ii) orthogonal M-ary frequency-shift keying modulated with non-coherent detection. The high signal-to-noise ratio approximations of the average SER are also obtained to investigate the effects of modulation order and relay placement on the system's performance.

Performance Evaluation of Visible Light Communication for DCO and ACO Optical OFDM Techniques

16:57 - 17:15

Mahendra Pratap Singh Bhadoria (IIT Delhi, India)

Gaurav Pandey (IIT Delhi, India)

Abhishek Dixit (IIT Delhi & IBBT, India)

Visible light communication (VLC) has evolved as a relatively new research topic that employs white light-emitting diodes for data transmission. The basic requirement for optical transmission is that signal should be real and positive. Among the various developed modulation techniques for VLC, orthogonal frequency division multiplexing (OFDM) has drawn major consideration because of its high data rate and heftiness to inter-symbol interference (ISI), but it suffers from the problem of high peak to average power ratio (PAPR), causing signal distortion thereby effecting the system efficiency. In this paper, asymmetrically clipped optical OFDM (ACO-OFDM) and direct current biased optical OFDM (DCO-OFDM) for VLC have been investigated by evaluating its bit error rate (BER) and PAPR performance. The effect of the uncorrelated noise due to dual-sided clipping of the signal, modulation order and number of subcarriers on the symbol distortion is also studied. Study on selection of optimum clipping levels is done to reduce the PAPR for both the schemes. Simulation analysis implies that ACO-OFDM is better than DCO-OFDM by around 4.5 dB for a BER of 1×10^{-3} and by about 2 dB for a PAPR complementary cumulative distribution function (CCDF) of 1×10^{-1} .

TECHNICAL PROGRAM

ML2: Machine Learning and Optimization

Friday, February 22, 15:45 - 17:15

Venue: MPA

Chair: Kunal Chaudhury (IISc, India)

Modelling and Short Term Forecasting of Flash Floods in an Urban Environment

15:45 - 16:03

Suraj Ogale (DAIICT, India)

Sanjay Srivastava (DAIICT, India)

Rapid urbanization, climate change, and extreme rainfall have resulted in a growing number of cases of urban flash floods. It is important to predict the occurrence of a flood so that the aftermath of it can be minimized. As the name suggests, an urban flash flood occurs in an urban area in a very short span of time. To reduce the impact of these events, short-term forecasting or nowcasting is used for prediction of the very near future incident. In orthodox methods of flood forecasting, current weather conditions are examined using conventional methods such as the use of radar, satellite imaging and calculations involving complicated mathematical equations. However, recent developments in Information and Communication Technology (ICT) and Machine Learning (ML) has helped us to study this hydrological problem from a different perspective. The aim of this paper is to design a theoretical model considering the parameters causing the urban flash flood and predict the event beforehand. To test the soundness model, data syntheses is performed and the results are checked using the artificial neural network.

An Iterative Eigensolver for Rank-Constrained Semidefinite Programming

16:03 -16:21

Rajat Sanyal (KPMG Advisory Services Private Limited, India)

Aditya Singh (IISc Bangalore, India)

Kunal Chaudhury (IISc Bangalore, India)

Rank-constrained semidefinite programming (SDP) arises naturally in various applications such as max-cut, angular (phase) synchronization, and rigid registration. Based on the alternating direction method of multipliers, we develop an iterative solver for this nonconvex form of SDP, where the dominant cost per iteration is the partial eigendecomposition of a symmetric matrix. We prove that if the iterates converge, then they do so to a KKT point of the SDP. In the context of rigid registration, we perform several numerical experiments to study the convergence behavior of the solver and its registration accuracy. As an application, we use the solver for wireless sensor network localization from range measurements. The resulting algorithm is shown to be competitive with existing optimization methods for sensor localization in terms of speed and accuracy.

TECHNICAL PROGRAM

A Weighted Optimization for Fourier Ptychographic Microscopy

16:21 - 16:39

Parimala Kancharla (IIT Hyderabad, India)

Sumohana Channappayya (IIT Hyderabad, India)

Fourier ptychography can be implemented as a phase retrieval optimization algorithm that iteratively solves for high resolution spectrum from low resolution images. In prior art, all the low resolution images were considered equally in the optimization. In this paper, we propose a weighted optimization algorithm to enhance the quality of reconstruction with the same convergence speed. Our method is motivated by the observation that bright field and dark field low resolution images have significantly different pixel intensities. Therefore, we weight their estimated error differently in the optimization. Though the proposed method is both conceptually and computationally simple, it dramatically improves the quality of reconstruction. We also show that the weighted optimization algorithm converges to a lower mean squared error value compared to the conventional optimization. We validate our approach on several low resolution images from an experimental dataset.

Top-m Clustering with a Noisy Oracle

16:39 - 16:57

Tuhinangshu Choudhury (IIT Bombay, India)

Dhruti Shah (IIT Bombay, India)

Nikhil Karamchandani (IIT Bombay, India)

In this paper, we analyse the problem of top-m clustering with access to a noisy oracle. We consider a model where there are n nodes, belonging to k clusters. We have access to an oracle which when queried with a pair of nodes, returns a binary answer indicating whether they belong to the same cluster or not, but with a probability of error p . Our goal is to identify the top-m clusters in terms of size, using the noisy answers from the oracle. This setting was recently studied in [9], which provides an iterative algorithm for the case of complete clustering, i.e., $m = k$. We identify conditions (on the relative sizes of clusters) under which the first m stages of the algorithm would recover the top m clusters. We also analyze the query complexity of the algorithm and provide an upper bound which is a function of the number of recovered clusters m and the sizes of the top clusters.

Unsupervised GIST Based Clustering for Object Localization

16:57 - 17:15

Saprem Shah (DAIICT, India)

Kunal Khatri (DAIICT, India)

Purva Mhasakar (DAIICT, India)

Rajendra Nagar (IIT Gandhinagar, India)

Shanmuganathan Raman (IIT Gandhinagar, India)

In the past years, there have been several attempts for the task of object localization in an image. However, most of the algorithms for object localization have been either supervised or weakly supervised. The work presented in this paper is based on the localization of a single object instance, in an image, in a fully unsupervised manner. Initially, from the input image, object proposals are generated where the proposal score for each of these proposals is calculated using a saliency map. Next, a graph by the GIST feature similarity between each pair of proposals is constructed. Density-based spatial clustering of applications with noise (DBSCAN) is used to make clusters of proposals based on GIST similarity, which eventually helps us in the final localization of the object. The setup is evaluated on two challenging benchmark datasets - PASCAL VOC 2007 dataset and object discovery dataset. The performance of the proposed approach is observed to be comparable with various state-of-the-art weakly supervised and unsupervised approaches for the problem of localization of an object.

TECHNICAL PROGRAM

COM4: Communication System Design

Saturday, February 23, 11:45 – 13:15

Venue: GJH

Chair: Sibi Raj B Pillai (IIT Bombay, India)

Development of an Efficient Low-complexity Channel Estimator for Digital Television Terrestrial Broadcasting Systems

11:45 - 12:03

Ghanshyamkumar Sah (IIT Roorkee, India)

Pyari Mohan Pradhan (IIT Roorkee, India)

Orthogonal frequency division multiplexing (OFDM) is widely used to transmit data in many wireless communication applications including digital television terrestrial broadcasting (DTTB). Although the existing dual pseudo noise padding (DPNP) based time-domain synchronous OFDM (TDS-OFDM) system has low complexity, the spectral efficiency is low. The time-frequency-domain (TFD) based frame structure enhances the system performance of TDS-OFDM over fast time-varying channels by compromising with computational complexity. This paper proposes a novel frame structure for OFDM-based DTTB system which incorporates pilots in the time domain, and retains the cyclic prefix and modulable orthogonal sequence (MOS) from the TFD-based frame structure. Since the proposed frame structure is completely defined in time domain, channel estimation and equalization become easier. Using the new frame structure, a novel channel estimation technique is proposed that works in two stages. In the first stage, the MOS in guard interval is used to estimate the channel delay and gain. In the second stage, the channel gains estimated in first stage are fine-tuned using adaptive algorithms such as least mean square (LMS) or recursive least squares algorithm. The bit-error-rate (BER) performance of the proposed two-stage channel estimation technique is better compared to that of DPNP-based TDS-OFDM. In addition, computational complexity of the proposed LMS-based two-stage channel estimation approach is low compared to TFD-based TDS-OFDM system. Less than 1.5% of the total sub-carriers are used as redundant pilots, and therefore the loss in spectral efficiency is negligible in the proposed approach.

Power Domain NOMA Design Based on MBER Criterion

12:03- 12:21

Amit Dutta (IIT Kharagpur, India)

Non-orthogonal multiple access (NOMA) has been gaining notable attention in the context of next generation communication system. The key primary benefit includes the higher spectrum efficiency compared to its various orthogonal counterparts. In this treatise, we consider a power NOMA design based on the minimum bit error ratio (MBER) criterion. Inspiration has been drawn from the MBER based works, which show a considerable performance improvement in terms of bit error ratio (BER) for a system. In this work, we have considered a single-input single-output (SISO) system with quadrature phase shift keying (QPSK) signal constellation. The numerical results demonstrate an overall BER improvement compared to the existing schemes, albeit, the scheme attracts large computational complexity. Traditionally, NOMA increases the spectral efficiency compared to its orthogonal counterpart. Nevertheless, our proposed solution will still hold this feature along with a better BER performance, though its spectral efficiency will be less compared to the traditional sum-rate based power NOMA.

TECHNICAL PROGRAM

Full-duplex Multi-user Pair Scheduling with Time-selective Fading and Imperfect CSI

12:21 – 12:39

Prabhat Kumar Sharma (VNIT, India)

Prasanna Raut (VNIT, India)

A multi-user full-duplex (FD) two-way communication system with decode-and-forward (DF) relaying protocol is investigated over time-selective fading channels. The effect of imperfect channel state information (CSI) is considered. The fading channel based approach is used to characterize the residual self-interference (RSI) at FD nodes. The outage performance of the considered system is investigated for different scheduling schemes based on the availability of CSI at the relay node. We derive the closed-form tight approximate expressions for the system outage probability assuming independent and non-identically distributed Rayleigh fading channels. Further, the tightness of the approximation presented is verified through Monte-Carlo simulations. Our analysis reveals the significant insights about the impact of time-selective fading, imperfect CSI, and RSI on the performance of the considered system.

A Planar Four-Port Integrated UWB and NB Antenna System for CR in 3.1GHz to 10.6GHz

12:39- 12:57

Anveshkumar Nella (VIT Bhopal, India)

Abhay Gandhi (VNIT, India)

This paper is aimed to present a planar four-port integrated UWB and narrowband (NB) antenna system for cognitive radio (CR) technology in the UWB 3.1GHz to 10.6GHz. This system consists of a UWB antenna for spectrum monitoring and three NB antennas for communication. These ultra wideband and narrowband antennas are incorporated on an FR-4 substrate having dimensions 28mmx31mmx1.6mm. The ultra wideband antenna, attached to port 1, is capable of monitoring the complete FCC unlicensed UWB spectrum 3.1GHz to 10.6GHz. The three NB antennas accomplish either single or dual bands to access the complete 3.1GHz to 10.6GHz band for communication. In particular, the first narrowband antenna, linked at port 2, attains a single band ranging from 8.26GHz to 11.16GHz. The second narrowband antenna, allied at port 3, also yields a single operating band ranges from 4.29GHz to 6GHz. The third NB antenna, associated with port 4, achieves a dual band behaviour starting from 3.06GHz to 4.49GHz and 5.97GHz to 8.35GHz. The coupling between the antennas is less than -17dB across the complete UWB. The proposed antenna system is fabricated and tested. It is observed that there is a good agreement between the simulated and measured results.

Generalized Selection Combining for Dynamic SSK-BPSK Systems

12:57- 13:15

Ananth A (IIITDM Kancheepuram, India)

Palani Maheswaran (IIT Madras, India)

Mandha Damodaran Selvaraj (IIITDM Kancheepuram, India)

Space shift keying (SSK) is a multiple-input multiple-output (MIMO) technique in which the transmitter can be designed with a single radio frequency (RF) chain. By adaptively selecting the modulation in a two antenna transmitter as either SSK or binary phase shift keying (BPSK), dynamic SSK-BPSK (DSB) obtains second order transmit diversity. In this work, we conceive DSB with generalized selection combining (DSB-GSC) to reduce the receiver circuit complexity. Specifically, we propose the metrics of modulation selection and receiver antenna selection for DSB where the receiver is equipped with lesser number of RF chains than its antennas. The performance of DSB-GSC is analyzed with exact bit error rate (BER) expression which is validated using simulation results. From the results, we infer that DSB-GSC provides diversity order equal to twice the number of receiver antennas irrespective of the number of RF chains used at the receiver. We further infer that there is only small SNR gains attained for increasing receiver RF chains. Thus the receiver complexity of the system can be considerably reduced with a small performance loss compared to that of full complex receiver.

TECHNICAL PROGRAM

SP3: Image Processing

Saturday, February 23, 11:45 – 13:15

Venue: MPA

Chair: Vinod Pankajakshan (IIT Roorkee, India)

Camera Zoom Detection and Classification Based on Application of Histogram Intersection and Kullback Leibler Divergence

11:45 - 12:03

Pavan Sandula (NIT Rourkela, India)

Manish Okade (NIT Rourkela, India)

This paper presents a novel compressed domain technique for detecting zooming camera in video sequences and its further classification into zoom-in camera and zoomout camera. The inter-frame block motion vector field serves as the input to the proposed system which is partitioned into four representative quadrants for analysis purposes. The histograms of these four quadrants are analyzed utilizing histogram intersection feature for zoom motion detection while the cumulative histogram of these four quadrants are analyzed utilizing Kullback-Leibler divergence feature for zoom motion classification purposes. Experimental validation carried out utilizing block motion vectors extracted using Exhaustive Search Motion Estimation algorithm as well as H.264 decoded block motion vectors demonstrate superior performance in comparison to existing techniques.

Splitting Merged Characters of Kannada Benchmark Dataset Using Simplified Paired-Valleys and L-Cut

12:03- 12:21

Shiva Kumar H. Ramagowda (IISc Bangalore, India)

Madhavaraj Ayyavu (IISc Bangalore, India)

Ramakrishnan Ganesan (IISc Bangalore, India)

We reduce the computational complexity of the paired-valley algorithm for splitting merged characters, from $O(N^2)$ down to $O(N)$, where N is the number of symbols merged. We also propose an effective way (L-cut algorithm) to separate the merged half-consonants (known in Kannada as *ottus*) from the base symbols. We have created a benchmark dataset of 4033 sub-word images in Kannada, each comprising two or more merged characters. We test the recognition accuracy of Tesseract OCR on the created benchmark dataset, before and after applying our technique. The accuracy of Tesseract v3 OCR on the created dataset of 61.6% increases by 20% to a value of 81.7% after the splitting of the characters by our method. The algorithm's scalability to other scripts has been explored by limited experiments on Telugu and Tamil.

TECHNICAL PROGRAM

Gamma Enhanced Binarization - An Adaptive Nonlinear Enhancement of Degraded Word Images for Improved Recognition of Split Characters

12:21 – 12:39

Shiva Kumar H. Ramagowda (IISc Bangalore, India)

Ramakrishnan Ganesan (IISc Bangalore, India)

Recognition performance of any OCR suffers because of the merged and split characters that occur in the scanned images of degraded printed documents. We propose an elegant method of non-linearly enhancing such degraded, gray-scale word images. This connects the broken strokes of the characters, so that binarization of the processed word images gives components with better connectivity for most characters or recognizable units. From an initial value of one, the value of gamma, the parameter determining the enhancement, is decreased in powers of 2 and the right value of gamma is chosen based on the recognition score of our character classifier. We have created a benchmark dataset of 1685 degraded word images obtained from scanned pages of several old Kannada books. The word images have been recognized before and after the proposed nonlinear enhancement. There is an absolute improvement of 14.8% in the Unicode level recognition accuracy of our SVM-based character classifier on the above dataset due to the proposed enhancement of the gray-scale word images. Even on the Google's Tesseract OCR for Kannada, our gamma enhanced binarization results in an improvement of 5.6% in the Unicode level accuracy.

Full Reference Stereoscopic Video Quality Assessment Based on Spatio-Depth Saliency and Motion Strength

12:39- 12:57

Sameeulla Khan Md (VIT Amaravati, India)

Sumohana Channappayya (IIT Hyderabad, India)

Stereoscopic video quality is a perceptual phenomena that is related to the human visual system (HVS). In this paper, we present a spatio-depth saliency and motion strength based full reference stereoscopic video quality metric (FRSVQA). Initially, we obtain a spatial distortion map on every video frame to estimate spatial quality. The spatial distortion map is then refined by the depth salient maps to estimate depth quality. We also estimate the temporal quality by refining the spatial distortion map with the inter-frame difference map at the locations specified by motion edges. The spatial, depth and temporal qualities are systematically combined and averaged over the frames to estimate the overall stereo video quality metric.

Interpolated Compressed Sensing for Calibrationless Parallel MRI Reconstruction

12:57- 13:15

Bhabesh Deka (Tezpur University, India)

Sumit Datta (Tezpur University, India)

Parallel magnetic resonance imaging (pMRI) in clinical study are commonly acquired in multiple slices; parallelly along different channels. Since, MRI traditionally suffers from slow data acquisition, reconstruction of images in clinical pMRI would be further slower. Compressed sensing MRI (CS-MRI) has successfully demonstrated its potential in reducing the scan time of pMRI by manifolds. Due to high correlation of adjacent slices in multislice sequence, interpolation of multi-slice data may be carried out to support non-uniform undersampling based CS reconstruction of slices in k-space. Exploiting intra/inter slice as well as multichannel data redundancy of multi-slice pMRI, it is possible to accelerate the scan time further. These correlations can be well modeled by introducing multidimensional wavelet forest sparsity and joint total variation regularization during the CS reconstruction. To validate our claim, a number of experiments are carried out with real pMRI datasets and results are compared with the state-of-the-art.

TECHNICAL PROGRAM

COM5: Information and Coding Theory

Saturday, February 23, 14:15 - 15:45

Venue: GJH

Chair: V. Lalitha (IIIT Hyderabad, India)

Performance Analysis of BCH and Repetition Codes in Gamma-Gamma Faded FSO Link

14:15 – 14:33

Sonali Garg (IIT Delhi, India)

Nancy Gupta (IIT Delhi, India)

Abhishek Dixit (IIT Delhi & IBBT, India)

V K Jain (IIT Delhi, India)

In this research paper, we have analyzed free space optical (FSO) link performance for intensity modulated/ direct detection system. FSO communication link is key to the next generation 5G/10G wireless networks. Generally, FSO link performance is impaired by atmospheric turbulence. Utilizing error correcting codes (ECCs) is one of the mitigation techniques employed to reduce the impact of atmospheric turbulence. We explore Bose–Chaudhuri–Hocquenghem (BCH) and repetition codes for mitigating the effects caused by atmospheric turbulence. We have made the performance comparison in terms of bit error rate (BER) under different turbulence regimes, viz., low, moderate and high regime for both uncoded and coded systems. We evaluate the analytical results and validate them with the simulations. It is concluded that BCH coded system performance is always better than the repetition coded system in all the turbulence regimes. BCH coded system provide a coding gain of 23.4 dB at a high turbulence level which reduces to 17.5 dB and 5 dB at moderate and low turbulence regimes, respectively. However, when we combine the benefits of both the coding schemes, the performance improvement obtained is much higher when compared with both the codes individually. It is, of course, at the cost of an increase in transmission bandwidth requirement.

Towards the Exact Rate Memory Tradeoff in Coded Caching

14:33 - 14:51

Vijith kumar K P (IIT Guwahati, India)

Brijesh Kumar Rai (IIT Guwahati, India)

Tony Jacob (IIT Guwahati, India)

Caching plays an important role in improving internet performance by keeping a fraction of the files closer to the end user. The peak data traffic in the network can be significantly reduced by proper utilization of caching. Recent studies have shown that coded caching does help in further reducing the data traffic over uncoded caching. In this paper, we consider the problem of the exact rate memory tradeoff in coded caching. For the (3, 3) canonical cache network, a new caching scheme to achieve the memory rate pair (5/3,1/2) is introduced. This scheme is further extended to the (4,4) canonical cache network, to achieve the memory rate pair (11/4,1/3). We prove the optimality of both the proposed schemes by deriving new lower bounds and thus partially characterizing the exact rate memory tradeoff in coded caching.

TECHNICAL PROGRAM

On the Optimality of Simple Han-Kobayashi Schemes for Gaussian Interference Channels

14:51-15:09

Ragini Chaluvadi (IIT Madras, India)

Srikrishna Bhashyam (IIT Madras, India)

The Generalized Degrees of Freedom (GDoF) region of the 2-user Gaussian Interference Channel (GIC) was derived by Etkin, Tse and Wang. This GDoF region is achieved using the class of Han-Kobayashi (HK) schemes. For K-user GICs with $K > 2$, the GDoF region is not known completely. For the K-user GIC, Geng et al. Derived the channel conditions under which Gaussian signalling and Treating Interference as Noise (TIN) is GDoF optimal. The TIN scheme also belongs to the class of HK schemes. In this paper, we derive conditions under which Simple HK (S-HK) schemes are GDoF optimal for general K-user GICs. Simple HK schemes are HK schemes with Gaussian signalling, no time sharing, and no private-common power splitting. The class of simple HK schemes includes the TIN scheme and schemes that involve various levels of interference decoding and cancellation at each receiver.

A Minimax Theorem for Finite Block length Joint Source-Channel Coding over an AVC

15:09- 15:27

Anuj Vora (IIT Bombay, India)

Ankur A. Kulkarni (IIT Bombay, India)

We pose the finite blocklength communication problem in the presence of a jammer as a zero-sum game between the encoder-decoder team and the jammer, where the communicators, as well as the jammer, are allowed locally randomized strategies. The minimax value of the game corresponds to joint sourcechannel coding over an Arbitrarily Varying Channel (AVC), which in the channel coding setting is known to admit a strong converse. The communicating team's problem is non-convex and hence, in general, a minimax theorem need not hold for this game. However, we show that an approximate minimax theorem holds in the sense that the minimax and maximin values of the game approach each other asymptotically. In particular, for rates above a critical threshold, both the minimax and maximin values approach unity. This result is stronger than the usual strong converse for channel coding over an AVC, which only says that the minimax value approaches unity for such rates.

General Compute and Forward for Virtual Full-Duplex Relaying

15:27- 15:45

Roshan Sam (IIT Madras, India)

Antony Mampilly (IIT Madras, India)

Srikrishna Bhashyam (IIT Madras, India)

Motivated by the wireless backhaul application, multihop virtual full duplex relaying using a successive relaying protocol based on compute-and-forward (CoF) was proposed recently by Hong and Caire. The channel gain in each hop was assumed to be equal. In this paper, we consider multihop virtual full duplex relaying where the gain in the different hops can be unequal. We use the recently proposed general compute-and forward (GCoF) scheme along with successive relaying. GCoF eliminates the non-integer penalty present in CoF or the CoF with simple power allocation used earlier. We determine the achievable rate of virtual full duplex relaying using GCoF for the multihop case and show that this rate is within a constant gap (also independent of the number of hops) of the cutset upper bound under some assumptions.

TECHNICAL PROGRAM

SP4: Biomedical Signal Processing

Saturday, February 23, 14:15-15:45

Venue: MPA

Chair: Sumohana Channappayya (IIT Hyderabad, India)

A SegNet Based Image Enhancement Technique for Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video

14:15 – 14:33

Renuka Mannem (IISc Bangalore, India)

Valliappan CA (IISc Bangalore, India)

Prasanta Kumar Ghosh (IISc Bangalore, India)

In this paper, we propose a new technique for segmentation of the Air-Tissue Boundaries (ATBs) in the upper airway of the vocal tract in the midsagittal plane of the realtime Magnetic Resonance Imaging (rtMRI) videos. The proposed technique uses a segmentation using Fisher-discriminant measure (SFDM) scheme. The paper introduces an image enhancement technique using semantic segmentation in the preprocessing of the rtMRI frames before ATB prediction. We use a deep convolutional encoder-decoder architecture (SegNet) for semantic segmentation of the rtMRI images. The paper examines the significance of the preprocessing before ATB prediction by implementing the SFDM approach with different preprocessing techniques. Experiments with 5779 rtMRI video frames from four subjects demonstrate that using the semantic segmentation based image enhancement of rtMRI frames, the performance of the SFDM approach is improved compared to the other preprocessing approaches. Experiment results also show that the proposed approach yields 8.6% less error in ATB prediction compared with a semi-supervised grid based baseline segmentation approach.

Edge Preserved Herringbone Artifact Removal from MRI Using Two-Stage Variational Mode Decomposition

14:33– 14:51

Divya Pankaj (Amrita University, Coimbatore, India)

Govind D (Amrita Vishwa Vidyapeetham, India)

NarayananKutty Kotheneth K a (Amrita School of Engineering, India)

Magnetic Resonance Imaging (MRI) is an efficient and non-invasive method for analyzing the structural features and functional behaviors of internal organs and tissues for medical diagnosis. The artifacts present in MRI mislead the diagnostic procedure. Herringbone artifact is a hardware artifact generated from the outlier in k-space measurement. In realtime MRI, the herringbone artifact has non-stationary noise characteristics. The non-stationary noise characteristics affect the high-frequency characteristics which in turn results in an improper estimation of structural details of the image in the processing stage. The objective of the present work is to exploit the properties of the variational mode decomposition (VMD) in reducing the effective herringbone noise at selected spectral regions (high-frequency regions in particular) of the given MRI data. In the present work, the given herringbone artifact affected image is subjected to VMD in two stages. The reconstructed image by removing the higher frequency VMD modes in two stages found to enhance the noisy MRI data. In the second stage of processing, the discarded higher frequency VMD mode in the first stage is further decomposed into component modes in order to preserve the high-frequency details. The enhanced image is later reconstructed by adding low-frequency modes obtained in both decomposition stages. The effectiveness of the proposed two stage VMD based enhancement is confirmed from the improved scores obtained from the non-reference quality measures such as Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) and Naturalness Image Quality Evaluator (NIQE).

TECHNICAL PROGRAM

A Weighted SVM Based Approach for Automatic Detection of Posterior Myocardial Infarction Using VCG Signals

14:51-15:09

Eedara Prabhakararao (IIT Guwahati, India)
Samarendra Dandapat (IIT Guwahati, India)

Myocardial infarction (MI), commonly known as heart attack is a life-threatening arrhythmia occurs due to insufficient oxygen supply to the heart tissues resulted from formation of clots in one or more coronary arteries. There is a growing interest among researchers for automatic detection of MI using computer algorithms. Based on the spatial location of damaged tissues MI is further categorized as anterior MI, septal MI, lateral MI, inferior MI and posterior MI. Among all, automatic detection of posterior MI (PMI) with standard 12-lead electrocardiogram (12-lead ECG) signal is challenging as it does not have monitoring electrodes posterior to human body. In this paper, we propose an automatic method for PMI detection using 3-lead vectorcardiogram (3-lead VCG) signal. The proposed approach exploits changes in electrical conduction properties of heart tissues during cardiac activity for healthy control (HC) and PMI subjects in three-dimensional (3D) space. To quantify these changes multiscale eigen features (MSEF) of subband matrices are used. Furthermore, we propose a cost sensitive weighted support vector machine (WSVM) classifier to combat class imbalance, which is a common problem in real-world disease data classification. The publicly available PhysioNet/PTBDB diagnostic database has been used to validate the proposed method by using a total of 1463 HC, and 148 PMI 4 sec 3-lead VCG signals. The best test accuracy of 96.69%, sensitivity of 80%, and geometric mean of 88.72% are achieved by WSVM classifier with radial basis function (RBF) kernel.

Brain Tumor Segmentation Using Discriminator Loss

15:09-15:27

Joydeep Das (IIT Roorkee, India)
Rashmin Patel (IIT Roorkee, India)
Vinod Pankajakshan (IIT Roorkee, India)

The emerging field of Computer Vision has found enormous applications in our day-to-day lives and Medical Image Processing is one of the most prominent fields among them. Brain Tumor Segmentation is an important and challenging task because of the variety in shapes, sizes and texture content of the various types of brain tumors. Specifically, MICCAI BraTS organizes Brain Tumor Segmentation challenge every year. Since the evolution of CNNs it has obtained state-of-the-art results in the majority of computer vision related tasks. On BraTS Challenge 2017, an ensemble average of various CNN models (EMMA) holds the state-of-the-art performance. In this paper, we have proposed a model inspired by the classic Generative Adversarial Network (GAN). The proposed network has two models namely, Generator or Segmentor which generates label map of the input image and a Discriminator which helps the Generator model for an optimum solution by taking into account both short as well as long-distance spatial correlations between pixels with the help of a novel multi-scale loss function. The proposed architecture has three GANs in a cascaded fashion, each for Whole Tumor, Tumor Core and Enhancing Tumor, where the former network helps in effective reduction of false positives for the later networks. Our method also employs a multi-scale loss function derived from intermediate layers of Discriminator rather than depending just on a final layer cross-entropy loss. A multi-scale loss function also reduces unnecessary smoothing on contours. The proposed method performed comparatively better than the state-of-the-art techniques, having Dice scores of 0.820, 0.874 and 0.783 for Enhancing Tumor, Whole Tumor and Tumor Core respectively.

TECHNICAL PROGRAM

Decision Support System for Liver Cancer Diagnosis Using Focus Features in NSCT Domain

15:27-15:45

B Lakshmi Priya (Pondicherry Engineering College, India)
Kaliyaperumal Jayanthi (University of Pondicherry, India)
Biju Pottakkat (JIPMER, India)
G Ramkumar (JIPMER, India)

Diagnosis of liver cancer by medical experts using imaging modalities is found to be sub-optimal as different lesions exhibit similar visual appearance in the spatial domain. Thus computer aided diagnostic tools play a significant role in providing a decision support system for radiologists to minimize the risk of false diagnosis. This paper proposes a different feature set using focus operators for classifying different classes of liver cancer. As computation of focus measure involves the local neighborhood of pixel, focus operator is believed to indirectly measure the intricate texture details of the image. This knowledge of focus operator is exploited in NSCT domain to capture the directional components as feature variables replacing the classic texture features. The results in terms of classification accuracy and kappa coefficient proclaim that the focus operators can be employed as feature variables for classification scenario as it outperforms the state-of-the art texture features.

COM6: Detection and Estimation

Saturday, February 23, 14:15 - 15:45

Venue: MPC

Chair: Mandha Damodaran Selvaraj (IIITDM Kancheepuram, India)

Detection and Estimation of Multiple DoA Targets with Single Snapshot Measurements

14:15 – 14:33

Rakshith Jagannath (IIT Madras, India)

In this paper, we explore the problems of detecting the number of narrow-band, far-field targets and estimating their corresponding directions of arrivals (DoAs) from single snapshot measurements. We use the principles of sparse signal recovery (SSR) for detection and estimation of multiple targets. In the SSR framework, the DoA estimation problem is grid based and can be posed as the lasso optimization problem. The corresponding DoA detection problem reduces to estimating the optimal regularization parameter (τ) of the lasso problem for achieving the required probability of correct detection (P_c). We propose finite sample and asymptotic test statistics for detecting the number of sources with the required P_c at moderate to high signal to noise ratios. Once the number of sources are detected, or equivalently the optimal τ^* is estimated, the corresponding DoAs can be estimated by solving the lasso with regularization parameter set to τ^* .

TECHNICAL PROGRAM

Signal Design and Detection Algorithms for Quick Detection Under False Alarm Rate Constraints

14:33– 14:51

Pavan Kumar Reddy (IIST Trivandrum, India)

Vineeth Bala Sukumaran (IIST Trivandrum, India)

In this paper, we consider the design of sequential detection algorithms for the low delay detection of a finite duration transient change signal from noisy observations of the signal under a false alarm rate constraint. Such design problems are motivated by the need to detect explicit control signals with low delay. We propose five heuristic detection algorithms that include algorithms that directly estimate the start time of the signals. In contrast to prior work, we also consider the case where the transient change signal can be apriori designed so as to optimize the detection delay as well as false alarm rate. Using simulations and numerical studies, we compare the average delay and false alarm rate performance of the above algorithms for different choices of the transient change signals.

Improved Data Fusion for Multi-Sensor Tracking Using a Reinforced Viterbi Algorithm

14:51-15:09

Rajarshi Biswas (IIT Bombay, India)

Akash Doshi (IIT Bombay, India)

Akanksha Bhatta (IIT Bombay, India)

Sibi Raj B Pillai (IIT Bombay, India)

Employing multiple wide aperture radars with partially overlapping coverage to accurately track moving objects is becoming increasingly popular. However, identifying a common track across the radars can be challenging when each radar sensor obtains multiple measurements from different targets in its field of view. The presence of clutter and spurious measurements further complicates this problem. Data association and target tracking in this context can benefit from the combined processing of the sensor measurements. We adapt the well known single sensor Viterbi Data Association (VDA) algorithm to exchange information between multiple sensors, thereby reinforcing the target tracking performance. The proposed multi-sensor data fusion algorithm is demonstrated to have vastly improved performance over conventional single sensor techniques.

Single Versus Multi-Source Discrimination in Birdcalls Using Zero-Frequency Filtering

15:09-15:27

Ragini Sinha (IIT Mandi, India)

Vivek Vadluri (IIT Mandi, India)

Ashish Arya (IIT Mandi, India)

Padmanabhan Rajan (IIT Mandi, India)

In the processing of bioacoustic recordings such as birdcalls, sometimes it is desirable to determine if a recording has one bird calling or has more than one. In this paper, we utilize the well-established zero-frequency filtering method, used for determining significant instants of excitation (also called epochs), for this task. By determining the average number of epochs per second, we are able to reliably discriminate birdcalls made by a single bird from those made by multiple birds. Experimental evaluation on three bioacoustic datasets confirms the reliability of the method. Species identification studies using deep neural network classifiers highlight the utility of the method.

TECHNICAL PROGRAM

An Objective Measure to Assess Musical Noise Using Connected Time-Frequency Regions

15:27-15:45

Ajey Saligrama (PESIT-BSC, India)

H. G. Ranjani (IISc Bangalore, India)

Muralishankar R (CMR Institute of Technology, India)

Shankar H N (CMR Institute of Technology, India)

In this work, we propose an objective measure to assess the amount of musical noise that results from speech enhancement algorithms. The algorithms can result in non smooth suppression of background noise which in turn translates to isolated regions of high energy, referred to as musical noise. We propose to identify such regions by combining time-frequency (TF) bins associated through connectivity along with additional properties of these regions such as area, aspect ratio and total energy. The objective measure proposed is based on density of such regions. The effectiveness of the proposed measure is studied by correlating it with subjective assessment of listeners using enhanced speech of various algorithms.

SPONSORS

Diamond



Platinum



Gold



Silver



Best paper
award



Technical
Co-Sponsor



Others



AUTHOR INDEX

A, Ananth	Generalized Selection Combining for Dynamic SSK-BPSK Systems
A, Jishnu	Fronthual and Timing Standards for 5G
Adiga, Nagaraj	On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition
	Speaking-Rate Adaptation of Automatic Speech Recognition System Through Fuzzy Classification Based Time-Scale Modification
Ahamed, Shaik Rafi	High Performance Multiplierless Serial Pipelined VLSI Architecture for Real-Valued FFT
Ahmad, Waquar	On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition
	Speaking-Rate Adaptation of Automatic Speech Recognition System Through Fuzzy Classification Based Time-Scale Modification
Aktar, Jinia	Slot Antenna Miniaturization Using Copper Coated Circular Dielectric Material
Anghan, Mehul	Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier
Aroor Dinesh, Dileep	Emotion Recognition from Varying Length Patterns of Speech Using CNN-based Segment-Level Pyramid Match Kernel Based SVMs
Arya, Ashish	Single Versus Multi-Source Discrimination in Birdcalls Using Zero-Frequency Filtering
Ayyavu, Madhavaraj	Data-pooling and Multi-Task Learning for Enhanced Performance of Speech Recognition Systems in Multiple Low Resourced Languages
	Splitting Merged Characters of Kannada Benchmark Dataset Using Simplified Paired-Valleys and L-Cut
B, Ganga Gowri	Improved Epoch Extraction from Speech Signals Using Wavelet Synchrosqueezed Transform
B, Thangapandian	A Graph Based Clustering and Preconditioning of V-MIMO Wireless Sensor Networks
B Pillai, Sibi Raj	Channels with Action Dependent States and Common Reconstructions
	Improved Data Fusion for Multi-Sensor Tracking Using a Reinforced Viterbi Algorithm
Babu, Prabhu	Codebook Based Precoding for Multiuser MIMO Broadcast Systems: An MM Approach
Bala Sukumaran, Vineeth	Signal Design and Detection Algorithms for Quick Detection Under False Alarm Rate Constraints
Behura, Sambit	Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling
Bettagere, Bharath	Solving a Distributed Stochastic Optimization Problem How Good is the Drift-Plus-Penalty Algorithm
Bhadoria, Mahendra	Performance Evaluation of Visible Light Communication for DCO and ACO Optical OFDM Techniques
Bhashyam, Srikrishna	General Compute and Forward for Virtual Full-Duplex Relaying
	On the Optimality of Simple Han-Kobayashi Schemes for Gaussian Interference Channels
Bhati, Saurabhchand	Instantaneous Frequency Features for Noise Robust Speech Recognition
Bhatia, Vimal	ASER Analysis of General Order Rectangular QAM for Dual-Hop NLOS UV Communication System
Bhatta, Akankhya	Improved Data Fusion for Multi-Sensor Tracking Using a Reinforced Viterbi Algorithm
Bhattacharjee, Ratnajit	Design of Discrete Frequency-Coding Waveforms Using Phase-Coded Linear Chirp for Multiuser and MIMO Radar Systems

AUTHOR INDEX

Biswas, Rajarshi	Improved Data Fusion for Multi-Sensor Tracking Using a Reinforced Viterbi Algorithm	Chetupalli, Srikanth Raj	Comparison of Low Dimension Speech Segment Embeddings Application to Speaker Diarization
Bramhendra, Koilakuntla	Instantaneous Frequency Features for Noise Robust Speech Recognition	Choudhury, Tuhinangshu	Top-m Clustering with a Noisy Oracle
Burugula, Shashank Dhar	Instantaneous Frequency Features for Noise Robust Speech Recognition	D, Govind	Edge Preserved Herringbone Artifact Removal from MRI Using Two-Stage Variational Mode Decomposition
CA, Valliappan	A SegNet Based Image Enhancement Technique for Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video		Improved Epoch Extraction from Speech Signals Using Wavelet Synchrosqueezed Transform
Chakka, Vijay Kumar	A Graph Based Clustering and Preconditioning of V-MIMO Wireless Sensor Networks	Dandapat, Samarendra	A Weighted SVM Based Approach for Automatic Detection of Posterior Myocardial Infarction Using VCG Signals
Chakraborty, Mrityunjoy	An Improved Multitasking Diffusion APA Based on Controlled Inter-Cluster Collaboration	Das, Joydeep	Brain Tumor Segmentation Using Discriminator Loss
Chaluvadi, Ragini	On the Optimality of Simple Han-Kobayashi Schemes for Gaussian Interference Channels	Datta, Sumit	Interpolated Compressed Sensing for Calibrationless Parallel MRI Reconstruction
Chandra, Prafulla	Improved Tail Bounds for Missing Mass and Confidence Intervals for Good-Turing Estimator	De, Kishalaya	Emotion Recognition from Varying Length Patterns of Speech Using CNN-based Segment-Level Pyramid Match Kernel Based SVMs
Channappayya, Sumohana	A Weighted Optimization for Fourier Ptychographic Microscopy	Deb, Debasish	Design of Discrete Frequency-Coding Waveforms Using Phase-Coded Linear Chirp for Multiuser and MIMO Radar Systems
	Full Reference Stereoscopic Video Quality Assessment Based on Spatio-Depth Saliency and Motion Strength	Deka, Abhash	Development of Assamese Text-to-Speech System Using Deep Neural Network
	Full-Reference Video Quality Assessment Using Deep 3D Convolutional Neural Networks	Deka, Bhabesh	Interpolated Compressed Sensing for Calibrationless Parallel MRI Reconstruction
Chatterjee, Avhishek	Qubits Through Queues: The Capacity of Channels with Waiting Time Dependent Errors	Dendi, Sathya Veera Reddy	Full-Reference Video Quality Assessment Using Deep 3D Convolutional Neural Networks
Chaudhury, Kunal	An Iterative Eigensolver for Rank-Constrained Semidefinite Programming		
	Saliency Guided Image Detail Enhancement		
Chauhan, Rahul	A Circular Fractal Antenna Array		
Chepuri, Sundeep Prabhakar	Sparse Sampling for Product Graphs and Tensors		

AUTHOR INDEX

Deshmukh, Amit	Analysis and Resonant Length Formulation of Dual Band Microstrip Antenna with Modified Ground Proximity Fed Broadband Equilateral Triangular Microstrip Antenna Using Parasitic Rectangular Patches Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics Slot Cut Modified Triangular Shape Microstrip Antenna for Circular Polarization
Deshmukh, Rohan	Design of Multiband Negative Permittivity Metamaterial Based on Interdigitated and Meander Line Resonator
Deshmukh, Sanjay	Analysis and Resonant Length Formulation of Dual Band Microstrip Antenna with Modified Ground Proximity Fed Broadband Equilateral Triangular Microstrip Antenna Using Parasitic Rectangular Patches Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics Slot Cut Modified Triangular Shape Microstrip Antenna for Circular Polarization
Deshmukh, Siddharth	Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling
Dey, Bikash	Authenticated Communication over Multiple Access Channels with Adversarial Users
Dhaka, Kalpana	Performance Analysis of Wireless Powered Decode-and-Forward Relay System
Dixit, Abhishek	Performance Analysis of BCH and Repetition Codes in Gamma-Gamma Faded FSO Link Performance Evaluation of Visible Light Communication for DCO and ACO Optical OFDM Techniques
Doshi, Akash	Improved Data Fusion for Multi-Sensor Tracking Using a Reinforced Viterbi Algorithm
Doshi, Akshay	Analysis and Resonant Length Formulation of Dual Band Microstrip Antenna with Modified Ground Proximity Fed Broadband Equilateral Triangular Microstrip Antenna Using Parasitic Rectangular Patches Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics Slot Cut Modified Triangular Shape Microstrip Antenna for Circular Polarization
Dubey, Ankit	Hardware Implementation of Filtered OFDM for BB-PLC Using Software Defined Radio
Dutta, Amit	Power Domain NOMA Design Based on MBER Criterion
Enukonda Venkata, Pothan	Interference Violation Probability Constrained Underlay Cognitive Massive MIMO Network Under Imperfect Channel Knowledge
Francis, Fredy	Simulation of Emission Wavelength of Quantum Dot Based Single Photon Sources
Gandhi, Abhay	A Planar Four-Port Integrated UWB and NB Antenna System for CR in 3.1GHz to 10.6GHz
Ganesan, Ramakrishnan	Data-pooling and Multi-Task Learning for Enhanced Performance of Speech Recognition Systems in Multiple Low Resourced Languages Gamma Enhanced Binarization - An Adaptive Nonlinear Enhancement of Degraded Word Images for Improved Recognition of Split Characters
	Splitting Merged Characters of Kannada Benchmark Dataset Using Simplified Paired-Valleys and L-Cut

AUTHOR INDEX

Ganti, Radha Krishna	Overview of the Indian 5G Testbed
Garg, Kamal	ASER Analysis of General Order Rectangular QAM for Dual-Hop NLOS UV Communication System
Garg, Parul	Outage Analysis of an Asymmetric Dual Hop PLC-VLC System for Indoor Broadcasting
Garg, Sonali	Performance Analysis of BCH and Repetition Codes in Gamma-Gamma Faded FSO Link
Garnaik, Sarmila	Detection of Vowels in Speech Signals Degraded by Speech-Like Noise
Gavaskar, Ruturaj	Saliency Guided Image Detail Enhancement
Ghatak, Gourab	Millimeterwave Selection Optimization for Sustaining the 5G Use Cases
Ghosh, Prasanta	A SegNet Based Image Enhancement Technique for Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video
Ghosh, Sanjay	Saliency Guided Image Detail Enhancement
Goli, Srikanth	A Graph Based Clustering and Preconditioning of V-MIMO Wireless Sensor Networks
Gopalakrishnan, Anand	Comparison of Low Dimension Speech Segment Embeddings Application to Speaker Diarization
Gulati, Nihir	Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier
Gunde, Kiran	Modified Generalised Quadrature Spatial Modulation
Gupta, Akash	Outage Analysis of an Asymmetric Dual Hop PLC-VLC System for Indoor Broadcasting
Gupta, Nancy	Performance Analysis of BCH and Repetition Codes in Gamma-Gamma Faded FSO Link
Gupta, Rama	Interference Violation Probability Constrained Underlay Cognitive Massive MIMO Network Under Imperfect Channel Knowledge
Gupta, Sanjeev	A Circular Fractal Antenna Array
Gupta, Shalabh	Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier
Gupta, Shikha	Emotion Recognition from Varying Length Patterns of Speech Using CNN-based Segment-Level Pyramid Match Kernel Based SVMs
H N, Shankar	An Objective Measure to Assess Musical Noise Using Connected Time-Frequency Regions
Haque, Sk.	Slot Antenna Miniaturization Using Copper Coated Circular Dielectric Material
Hari, K. v. s.	Modified Generalised Quadrature Spatial Modulation
Hazarika, Jinti	High Performance Multiplierless Serial Pipelined VLSI Architecture for Real-Valued FFT
Hegde, Parikshit	Conditions for Optimality of Superposition Coding in Discrete Memoryless Broadcast Channels
Hiremath, Shrishail	Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling
Hota, Ashish	Game-Theoretic Vaccination Against Networked SIS Epidemics and Impacts of Human Decision-Making
Husain, Moula	Multimodal Fusion of Speech and Text Using Semi-supervised LDA for Indexing Lecture Videos
Jacob, Tony	Towards the Exact Rate Memory Tradeoff in Coded Caching
Jagadeesh, Harshan	Adversarial Attacks on Next-Generation Wireless Networks by Cognitive Radios
Jagannath, Rakshith	Detection and Estimation of Multiple DoA Targets with Single Snapshot Measurements

AUTHOR INDEX

Jagannatham, Aditya	Sparse Bayesian Learning (SBL)-Based Frequency-Selective Channel Estimation for Millimeter Wave Hybrid MIMO Systems	Kalyani, Sheetal	Tuning Free Algorithms for Sparse Estimation
Jagannathan, Krishna	Qubits Through Queues: The Capacity of Channels with Waiting Time Dependent Errors	Kamran, Rashmi	Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier
Jain, Sambhav	Caching Partial Files for Content Delivery	Kancharla, Parimala	A Weighted Optimization for Fourier Ptychographic Microscopy
Jain, V	Analysis of Beam Wander Effect of Flat-topped Multi-Gaussian Beam for FSO Communication Link	Kanth, Dhulipudi	Performance Analysis and Optimization of Interference Limited Multi-Antenna BRN
	Performance Analysis of BCH and Repetition Codes in Gamma-Gamma Faded FSO Link	Kar, Subrat	Analysis of Beam Wander Effect of Flat-topped Multi-Gaussian Beam for FSO Communication Link
Jani, Manan	Outage Analysis of an Asymmetric Dual Hop PLC-VLC System for Indoor Broadcasting		Analysis of Mid-Haul Characteristics for LTE-NR Multi-Connectivity in Heterogeneous Cloud RAN
Jayanthi, Kaliyaperumal	Decision Support System for Liver Cancer Diagnosis Using Focus Features in NSCT Domain	Karamchandani, Nikhil	Top-m Clustering with a Noisy Oracle
Jha, Pranav	Unified Control of Multi-RAT Radio Access Network: An SDN & NFV based Approach	Kashyap, Navin	Achieving Secrecy Capacity of Minimum Storage Regenerating Codes for All Feasible (n,k,d) Parameter Values
Jhunjhunwala, Ashok	Making Difficult Things Doable by Leveraging Communications A Case Study of Electric Vehicles in India	Kashyap, Salil	Interference Violation Probability Constrained Underlay Cognitive Massive MIMO Network Under Imperfect Channel Knowledge
K, Samudravijaya	Development of Assamese Text-to-Speech System Using Deep Neural Network	Kathania, Hemant	On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition
K a, Narayanankutty	Edge Preserved Herringbone Artifact Removal from MRI Using Two-Stage Variational Mode Decomposition		Speaking-Rate Adaptation of Automatic Speech Recognition System Through Fuzzy Classification Based Time-Scale Modification
K Manickam, Shivkumar	Probability Mass Functions for Which Sources Have the Maximum Minimum Expected Length	Kedia, Subham	Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling
K p, Vijith kumar	Towards the Exact Rate Memory Tradeoff in Coded Caching	Khan, Enamul	Slot Antenna Miniaturization Using Copper Coated Circular Dielectric Material
Kadam, Ameya	Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics	Khan, Imtiyaz	Performance Analysis and Optimization of Interference Limited Multi-Antenna BRN
Kadam, Poonam	Analysis and Resonant Length Formulation of Dual Band Microstrip Antenna with Modified Ground		

AUTHOR INDEX

Khan, Tasleem	High Performance Multiplierless Serial Pipelined VLSI Architecture for Real-Valued FFT	Lakshmi Priya, S	Improved Epoch Extraction from Speech Signals Using Wavelet Synchrosqueezed Transform
Khatri, Kunal	Unsupervised GIST Based Clustering for Object Localization	Lalitha, V.	Maximally Recoverable Codes with Hierarchical Locality
Kodukula, Sri Rama Murty	Instantaneous Frequency Features for Noise Robust Speech Recognition	M, Jyothish	Simulation of Emission Wavelength of Quantum Dot Based Single Photon Sources
Koonampilli J, Babu Narayanan	CeWiT's 5G Testbed Efforts	M, Meena	Multimodal Fusion of Speech and Text Using Semi-supervised LDA for Indexing Lecture Videos
KP, Sumesh	Hardware Implementation of Filtered OFDM for BB-PLC Using Software Defined Radio	Maheswaran, Palani	Generalized Selection Combining for Dynamic SSK-BPSK Systems
Krishnamurthy, Harikumar	Conditions for Optimality of Superposition Coding in Discrete Memoryless Broadcast Channels	Mampilly, Antony	General Compute and Forward for Virtual Full-Duplex Relaying
Krishnappa, Gokul	Full-Reference Video Quality Assessment Using Deep 3D Convolutional Neural Networks	Mandayam, Prabha	Differential Phase Encoding Schemes for Measurement-Device-Independent Quantum Key Distribution
Krishnaswamy, Dilip	Distributed Smart Networks: A convergence of 5G, IoT, AI, and Blockchain		Qubits Through Queues: The Capacity of Channels with Waiting Time Dependent Errors
Kulat, Kishore	Design of Multiband Negative Permittivity Metamaterial Based on Interdigitated and Meander Line Resonator	Mannem, Renuka	A SegNet Based Image Enhancement Technique for Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video
Kulkarni, Ankur	A Minimax Theorem for Finite Blocklength Joint Source-Channel Coding over an AVC	Marathe, Dushyant	Design of Multiband Negative Permittivity Metamaterial Based on Interdigitated and Meander Line Resonator
Kumar, Animesh	Learning a Bandlimited Field from Samples Taken at Unknown-Locations on a Path	MD, Sameeulla	Full Reference Stereoscopic Video Quality Assessment Based on Spatio-Depth Saliency and Motion Strength
Kumar, Avinash	Detection of Vowels in Speech Signals Degraded by Speech-Like Noise	Mhasakar, Purva	Unsupervised GIST Based Clustering for Object Localization
Kumar, Pawan	Performance Analysis of Wireless Powered Decode-and-Forward Relay System	Mishra, Jagabandhu	Modelling Glottal Flow Derivative Signal for Detection of Replay Speech Samples
Kumar Reddy, Pavan	Signal Design and Detection Algorithms for Quick Detection Under False Alarm Rate Constraints	Misra, Rajiv	Truthful Double Auction Based VM Allocation for Revenue-Energy Trade-Off in Cloud Data Centers
Lakshmi Priya, B	Decision Support System for Liver Cancer Diagnosis Using Focus Features in NSCT Domain	Moharir, Sharayu	Caching Partial Files for Content Delivery

AUTHOR INDEX

Mukherjee, Arka	Analysis of Beam Wander Effect of Flat-topped Multi-Gaussian Beam for FSO Communication Link	Pandey, Gaurav	Performance Evaluation of Visible Light Communication for DCO and ACO Optical OFDM Techniques
Mundlamuri, Rakesh	A Graph Based Clustering and Preconditioning of V-MIMO Wireless Sensor Networks	Panigrahi, Trilochan	Hardware Implementation of Filtered OFDM for BB-PLC Using Software Defined Radio
Murthy, Chandra	Codebook Based Precoding for Multiuser MIMO Broadcast Systems: An MM Approach	Pankaj, Divya	Edge Preserved Herringbone Artifact Removal from MRI Using Two-Stage Variational Mode Decomposition
Nagar, Rajendra	Unsupervised GIST Based Clustering for Object Localization	Pankajakshan, Vinod	Brain Tumor Segmentation Using Discriminator Loss
Nair, Aaditya	Maximally Recoverable Codes with Hierarchical Locality	Parvez, Khan	Slot Antenna Miniaturization Using Copper Coated Circular Dielectric Material
Nambath, Nandakumar	Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier	Patel, Diptiben	Adaptive Multiple-pixel Wide Seam Carving
Nayak, Shekhar	Ins	Patel, Rashmin	Brain Tumor Segmentation Using Discriminator Loss
	Error Performance of QAM GFDM Waveform with CFO Under AWGN and TWDP Fading Channel	Patel, Yashwant	Truthful Double Auction Based VM Allocation for Revenue-Energy Trade-Off in Cloud Data Centers
Nella, Anveshkumar	A Planar Four-Port Integrated UWB and NB Antenna System for CR in 3.1GHz to 10.6GHz	Pati, Debadatta	Modelling Glottal Flow Derivative Signal for Detection of Replay Speech Samples
Nemade, Harshal	Design of Discrete Frequency-Coding Waveforms Using Phase-Coded Linear Chirp for Multiuser and MIMO Radar Systems	Patra, Sarat	Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling
	High Performance Multiplierless Serial Pipelined VLSI Architecture for Real-Valued FFT	Patro, Ch Suraj	Sparse Bayesian Learning (SBL)-Based Frequency-Selective Channel Estimation for Millimeter Wave Hybrid MIMO Systems
Nighojkar, Animesh	Truthful Double Auction Based VM Allocation for Revenue-Energy Trade-Off in Cloud Data Centers	Phartiyal, Deepmala	LSTM-Deep Neural Networks Based Predistortion Linearizer for High Power Amplifiers
Ogale, Suraj	Modelling and Short Term Forecasting of Flash Floods in an Urban Environment	Pottakkat, Biju	Decision Support System for Liver Cancer Diagnosis Using Focus Features in NSCT Domain
Okade, Manish	Camera Zoom Detection and Classification Based on Application of Histogram Intersection and Kullback Leibler Divergence	Prabavathi, P	Design of Frequency-Signature Based Multiresonators Using Quarter Wavelength Open Ended Stub for Chipless RFID Tag
P, Soman	Improved Epoch Extraction from Speech Signals Using Wavelet Synchrosqueezed Transform		

AUTHOR INDEX

Prabhakar, Anil	Differential Phase Encoding Schemes for Measurement-Device-Independent Quantum Key Distribution	R, Sreenath	Dynamic Mode Selection and Link Adaptation in 5G NR
	Quantum Random Number Generator with One and Two Entropy Sources	Rai, Brijesh	Towards the Exact Rate Memory Tradeoff in Coded Caching
Prabhakaran, Vinod	Channels with Action Dependent States and Common Reconstructions	Rajan, Padmanabhan	Single Versus Multi-Source Discrimination in Birdcalls Using Zero-Frequency Filtering
Prabhakararao, Eedara	A Weighted SVM Based Approach for Automatic Detection of Posterior Myocardial Infarction Using VCG Signals	Ramachandran, Viswanathan	Channels with Action Dependent States and Common Reconstructions
Prabhu, Shailesh	Optical Wireless Communication (OWC) Using LiFi Based System and Optical Camera Communication (OCC) Based System	Ramagowda, Shiva Kumar	Gamma Enhanced Binarization - An Adaptive Nonlinear Enhancement of Degraded Word Images for Improved Recognition of Split Characters
Pradeep, Aditya	Improved Tail Bounds for Missing Mass and Confidence Intervals for Good-Turing Estimator		Splitting Merged Characters of Kannada Benchmark Dataset Using Simplified Paired-Valleys and L-Cut
Pradhan, Gayadhar	Detection of Vowel-Like Speech Using Variance of Sample Magnitudes	Ramalingam, C	Efficient Methods for Estimating Sinusoidal Frequencies Using Line Spectral Pairs
	Detection of Vowels in Speech Signals Degraded by Speech-Like Noise	Raman,	Adaptive Multiple-pixel Wide Seam Carving
Pradhan, Pyari	Development of an Efficient Low-complexity Channel Estimator for Digital Television Terrestrial Broadcasting Systems		Unsupervised GIST Based Clustering for Object Localization
Prakriya, Shankar	Towards Self-Sustaining Devices Through Energy Harvesting	Rameshwar, V. Arvind	Achieving Secrecy Capacity of Minimum Storage Regenerating Codes for All Feasible (n,k,d) Parameter Values
Prasanna, Mahadeva	Development of Assamese Text-to-Speech System Using Deep Neural Network	Ramkumar, G	Decision Support System for Liver Cancer Diagnosis Using Focus Features in NSCT Domain
	Modelling Glottal Flow Derivative Signal for Detection of Replay Speech Samples	Ranjani, H. g.	An Objective Measure to Assess Musical Noise Using Connected Time-Frequency Regions
Puli, Kishore	Detection of Vowel-Like Speech Using Variance of Sample Magnitudes	Ranu, Shashank	Differential Phase Encoding Schemes for Measurement-Device-Independent Quantum Key Distribution
R, Manivasakan	Simulation of Emission Wavelength of Quantum Dot Based Single Photon Sources	Raut, Prasanna	Full-duplex Multi-user Pair Scheduling with Time-selective Fading and Imperfect CSI
R, Muralishankar	An Objective Measure to Assess Musical Noise Using Connected Time-Frequency Regions		

AUTHOR INDEX

Rawat, Meenakshi	LSTM-Deep Neural Networks Based Predistortion Linearizer for High Power Amplifiers	Sanyal, Rajat	An Iterative Eigensolver for Rank-Constrained Semidefinite Programming
	Predistortion Linearizer Design for Ku Band RF Power Amplifier	Sarmah, Priyankoo	Development of Assamese Text-to-Speech System Using Deep Neural Network
Ray, Kamala Prasan	Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics	Selvamuthu, Dharmaraja	Analysis of Mid-Haul Characteristics for LTE-NR Multi-Connectivity in Heterogeneous Cloud RAN
	Slot Cut Modified Triangular Shape Microstrip Antenna for Circular Polarization	Selvaraj, Mandha Damodaran	Generalized Selection Combining for Dynamic SSK-BPSK Systems
Roy, Arijit	Design of Discrete Frequency-Coding Waveforms Using Phase-Coded Linear Chirp for Multiuser and MIMO Radar Systems	Shah, Dhruti	Top-m Clustering with a Noisy Oracle
S, Akarsh	Improved Epoch Extraction from Speech Signals Using Wavelet Synchrosqueezed Transform	Shah, Saprem	Unsupervised GIST Based Clustering for Object Localization
S, Ramakrishnan	Analysis of Mid-Haul Characteristics for LTE-NR Multi-Connectivity in Heterogeneous Cloud RAN		Detection of Vowels in Speech Signals Degraded by Speech-Like Noise
S, Subha Rani	Design of Frequency-Signature Based Multiresonators Using Quarter Wavelength Open Ended Stub for Chipless RFID Tag		On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition
Sah,	Development of an Efficient Low-complexity Channel Estimator for Digital Television Terrestrial Broadcasting Systems		Speaking-Rate Adaptation of Automatic Speech Recognition System Through Fuzzy Classification Based Time-Scale Modification
Sahu, Hemanta	SSK Performance with SWIPT Based Dual-Hop AF Relay over Rayleigh Fading		Adaptive Multiple-pixel Wide Seam Carving
Sahu, Pravas Ranjan	Error Performance of QAM GFDM Waveform with CFO Under AWGN and TWDP Fading Channel	Sharma, Govind	Sparse Bayesian Learning (SBL)-Based Frequency-Selective Channel Estimation for Millimeter Wave Hybrid MIMO Systems
	SSK Performance with SWIPT Based Dual-Hop AF Relay over Rayleigh Fading	Sharma, Prabhat	Full-duplex Multi-user Pair Scheduling with Time-selective Fading and Imperfect CSI
Saligrama, Ajey	An Objective Measure to Assess Musical Noise Using Connected Time-Frequency Regions	Shaw, Gautam	Quantum Random Number Generator with One and Two Entropy Sources
Sam, Roshan	General Compute and Forward for Virtual Full-Duplex Relaying	Singh, Aditya	An Iterative Eigensolver for Rank-Constrained Semidefinite Programming
Sandula, Pavan	Camera Zoom Detection and Classification Based on Application of Histogram Intersection and Kullback Leibler Divergence	Singh, Poonam	Performance Analysis and Optimization of Interference Limited Multi-Antenna BRN

AUTHOR INDEX

Singh, Prem	MIMO-FBMC Channel Estimation with Limited, and Imperfect Knowledge of Channel Correlations	Thoota, Sai Subramanyam	Codebook Based Precoding for Multiuser MIMO Broadcast Systems: An MM Approach
Singya, Praveen	ASER Analysis of General Order Rectangular QAM for Dual-Hop NLOS UV Communication System	Tripathi, Girish	Predistortion Linearizer Design for Ku Band RF Power Amplifier
Sinha, Ragini	Single Versus Multi-Source Discrimination in Birdcalls Using Zero-Frequency Filtering	V S Ch, Lakshmi Narayana	Caching Partial Files for Content Delivery
SR, Sivaram	Quantum Random Number Generator with One and Two Entropy Sources	Vadluri, Vivek	Single Versus Multi-Source Discrimination in Birdcalls Using Zero-Frequency Filtering
Sreenivas, Thippur	Comparison of Low Dimension Speech Segment Embeddings Application to Speaker Diarization	Vasudevan, Kasturi	MIMO-FBMC Channel Estimation with Limited, and Imperfect Knowledge of Channel Correlations
Srinivas, Nagapuri	Detection of Vowel-Like Speech Using Variance of Sample Magnitudes	Vaze, Rahul	Not Just Age but Age and Quality of Information
Srivastava, Sanjay	Modelling and Short Term Forecasting of Flash Floods in an Urban Environment	Vishnu, Palakkal	Efficient Methods for Estimating Sinusoidal Frequencies Using Line Spectral Pairs
Srivastava, Suraj	Sparse Bayesian Learning (SBL)-Based Frequency-Selective Channel Estimation for Millimeter Wave Hybrid MIMO Systems	Vora, Anuj	A Minimax Theorem for Finite Blocklength Joint Source-Channel Coding over an AVC
Tamarapalli, Venkatesh	The HTTP/2 Server Push and Its Implications on Mobile Web Quality of Experience	Yadav, Ishwar	Detection of Vowels in Speech Signals Degraded by Speech-Like Noise
Tamma, Bheemarjuna Reddy	An OAI Based Testbed for Cellular-WiFi Convergence		
Tarun Sai, Bandarupalli	Speaking-Rate Adaptation of Automatic Speech Recognition System Through Fuzzy Classification Based Time-Scale Modification		
Thangaraj, Andrew	Conditions for Optimality of Superposition Coding in Discrete Memoryless Broadcast Channels		
	Improved Tail Bounds for Missing Mass and Confidence Intervals for Good-Turing Estimator		
Thenkanidiyoor, Veena	Emotion Recognition from Varying Length Patterns of Speech Using CNN-based Segment-Level Pyramid Match Kernel Based SVMs		

CONTACT US / MAP

Neleesh B. Mehta

Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012
Email: ncc.iisc.2019@gmail.com

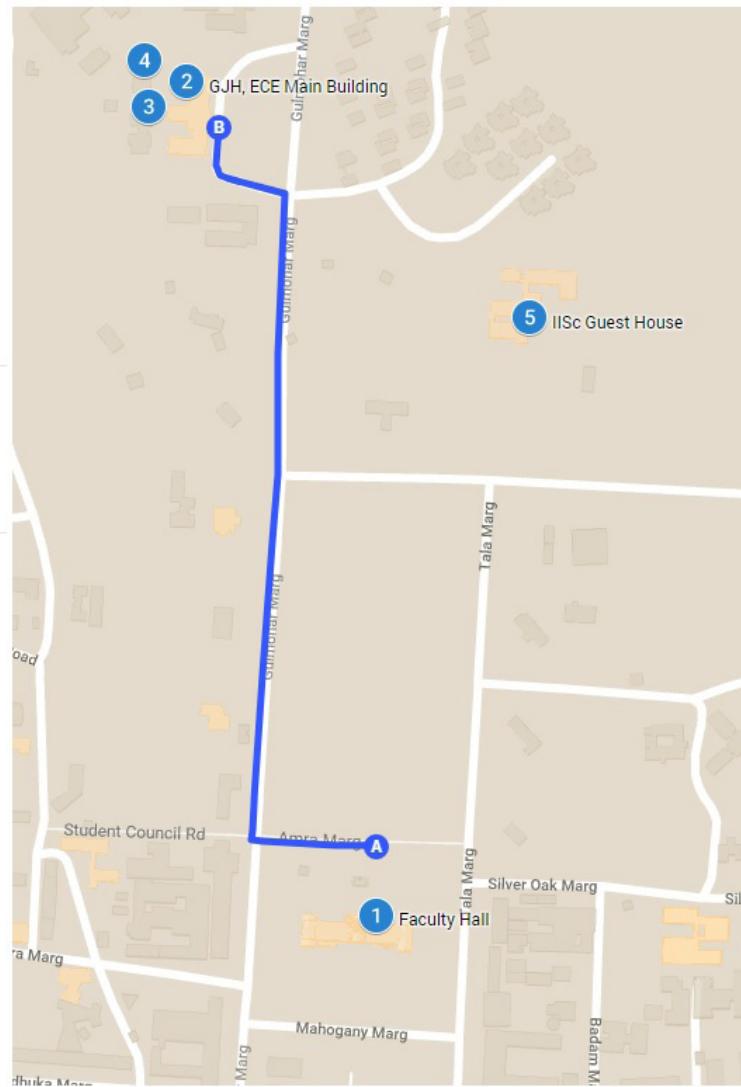
IISc Map -1

NCC Main Locations

- ① Faculty Hall
- ② GJH, ECE Main Building
- ③ MP Building (MPA, MPC)
- ④ Food Court
- ⑤ IISc Guest House

Directions from Faculty Hall to GJH, ECE Main Building

- A Faculty Hall
- B GJH, ECE Main Building



CONTACT US / MAP

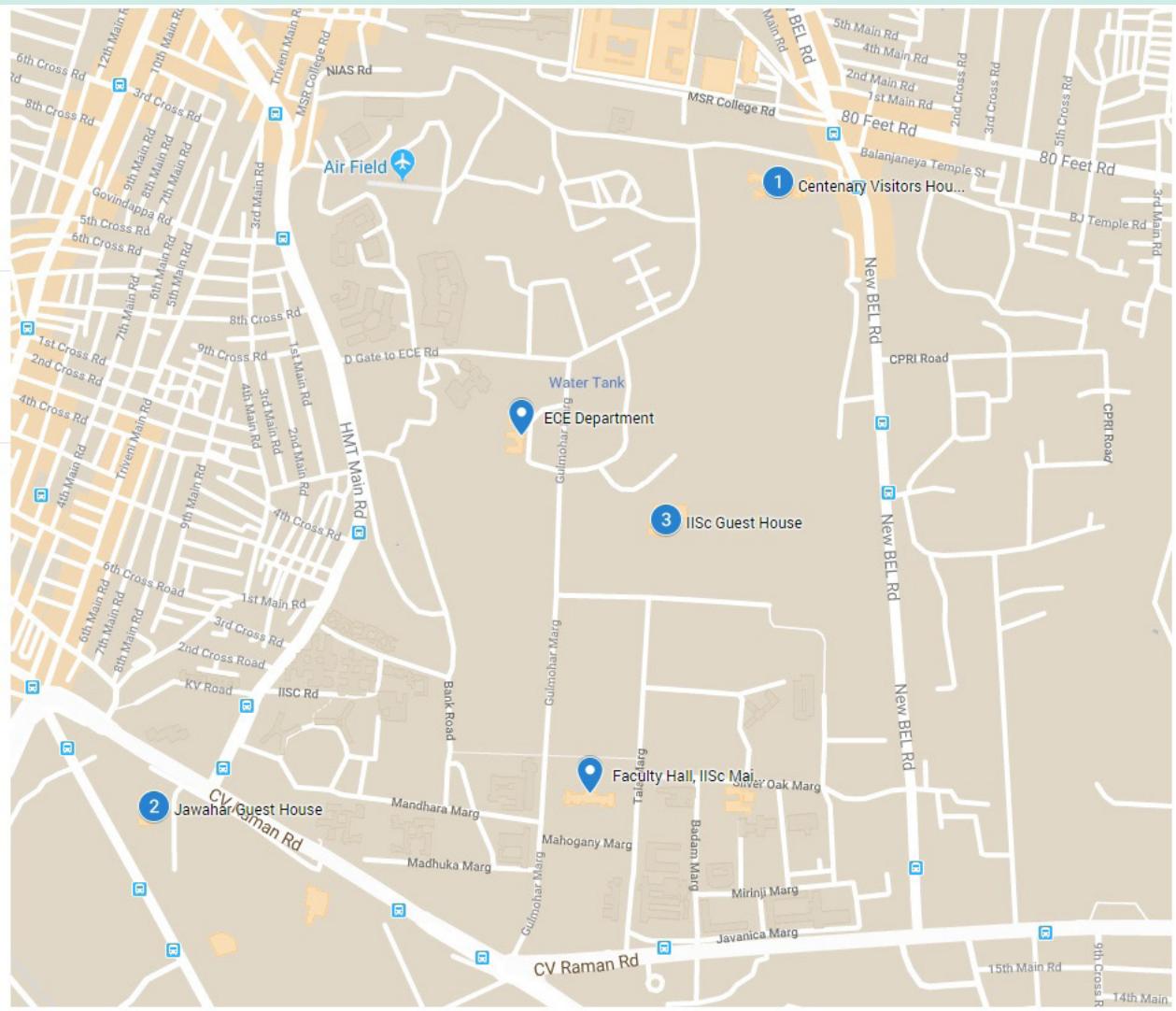
IISc Map - 2

NCC Guest House Locations

- 1 Centenary Visitors House
 - 2 Jawahar Guest House
 - 3 IISc Guest House

NCC Main Locations

-



Channels with Action Dependent States and Common Reconstructions

Viswanathan Ramachandran and Sibi Raj B Pillai

Department of Electrical Engineering

Indian Institute of Technology Bombay

{viswanathan,bsraj}@ee.iitb.ac.in

Vinod M. Prabhakaran

School of Technology and Computer Science

Tata Institute of Fundamental Research, Mumbai

vinodmp@tifr.res.in

Abstract—In channels with action dependent states, a common message is conveyed using two encoders operating sequentially, *viz.* an action encoder and a channel encoder. The actions drive the output of a discrete-memoryless channel (DMC), which in turn forms the state process for the DMC between the channel encoder and receiver. Assuming non-causal knowledge of the state-process at the channel encoder, a single letter characterization of the capacity is known in the discrete memoryless case.

We consider the action dependent state channel with a common message and an additional private message at the channel encoder, along with common reconstructions (CR) of the state process at the channel encoder and the decoder. Capacity characterizations for the discrete memoryless and Gaussian versions are presented. As a consequence, we settle the capacity characterization of the Gaussian action dependent channel with only a common message and CR. Moreover, by identifying a connection to degraded message sets multiple access channel (deg-MAC) models studied in literature, we establish the capacity regions for the discrete and Gaussian versions of deg-MAC with CR constraints.

I. INTRODUCTION

The problem of coding for channels with state was introduced in the seminal paper by Shannon [1], wherein causal state information was assumed at the encoder. The capacity for the non-causal setting was established by Gelfand and Pinsker [2]. In these models, the state process is assumed to be given by nature. Motivated by applications involving multi-stage encoding (for instance, two-stage recording on a magnetic storage device), Weissman [3] introduced the notion of a *channel with action dependent states (ADSC)*. In this setting, the transmitter can take *actions* that influence the formation of channel states in the first stage, and the encoding in the second stage is based upon the channel state sequence so generated and the message. Notice that the actions play a dual role of message communication as well as controlling the channel states. The capacity of this model was derived in [3].

The ADSC has close connections to the state-dependent degraded message sets multiple access channel (MAC) models of Somekh-Baruch *et al.* [4] and Zaidi *et al.* [5], [6]. In particular, the capacity for the discrete memoryless (DM) and Gaussian versions of the ADSC can be inferred from the results in [4]–[6]. Following [3], the action-dependent framework has been extended in several directions. Ahmadi *et al.* [7] studied

a discrete memoryless ADSC (with only a common message between the channel and action encoders) with the additional constraint of common reconstructions (CR) of the state at the encoder and decoder. Note that in a CR framework [8], the channel encoder and decoder must agree on an identical version of the state process. The capacity-distortion trade-off for the ADSC with CR in the Gaussian setting is hitherto unknown. In addition, some of the open problems which cannot be inferred from [3]–[7] are

- Discrete Memoryless Action Dependent State Channel with CR and additional private messages (PM) at the channel encoder,
- Gaussian Action Dependent State Channel with CR and private messages,
- State-Dependent Multiple Access Channels with Degraded Message Sets (deg-MAC) and CR.

We resolve these problems in the current paper. Notice that this also settles the optimality of Gaussian auxiliaries for the common message Gaussian ADSC with CR constraints, for which [7] gave only an achievable region. Table I summarizes the contributions of the current paper (checkmarked items) w.r.t. models already studied in literature.

Case	deg-MAC	ADSC	MAC+CR	ADSC+CR	ADSC+CR+PM
DMC	[5]	[3]	✓	[7]	✓
Gauss.	[5]	[5]	✓	✓	✓

TABLE I
SUMMARY OF CONTRIBUTIONS

Some other related literature on action-dependent models include Permuter *et al.* [9], wherein the source coding dual in which the decoder can take actions based on the observed compression index was addressed. Asnani *et al.* [10] considered a setting in which the encoder as well as the decoder can take probing actions to learn the channel state, with a cost constraint associated with each. Choudhuri *et al.* [11] considered causal state communication over an action-dependent channel and characterized the trade-off between message communication and state estimation distortion. Recently, Kittichokechai *et al.* [12] studied source and channel coding settings with action-dependent states and *reversible input* constraints, wherein input to the channel must be reconstructed reliably at the receiver. The action dependent model has been extended to multi-user channels as well— for instance see Steinberg *et al.* [13], [14].

In the sequel, we also note that in the absence of any state reconstruction constraints, our problem framework is identical to the co-operative multiple access model studied in [4] and [5]. The action encoder in our setting can be viewed to be playing the role of the non-cognizant encoder. For a proof of equivalence between the two settings, see Section III. From this equivalence, it follows that the capacity region of a Gaussian ADSC without CR is the same as that established in Theorem 4 of [5]. The novelty in our model is that we are also interested in CR of the state process in the Gaussian ADSC with additional private messages, which in turn allows us to establish the optimality of Gaussian auxiliaries for the model studied in [7]. The established equivalence is also used to determine the capacity region with CR constraints in the setting of [4], [5].

Organization: We introduce the system model and main results in Section II. The connection between our setting and degraded message sets MAC models without CR is established in Section III. The extension of our results to the Gaussian case is given in Section IV. Section V concludes the paper.

II. SYSTEM MODEL

Consider the model shown in Fig. 1. There are two encoders,

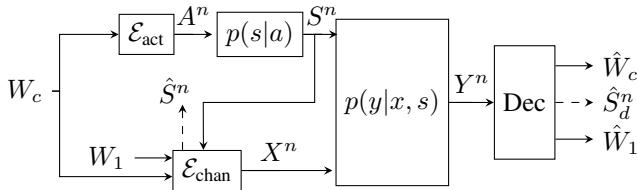


Fig. 1. Action-Dependent Channel with additional private message, and with/without common reconstructions

namely an action encoder \mathcal{E}_{act} and a channel encoder $\mathcal{E}_{\text{chan}}$. A common message W_c is observed by both \mathcal{E}_{act} and $\mathcal{E}_{\text{chan}}$. The action encoder chooses the output symbol $a \in \mathcal{A}$ which is fed to a DMC $p(s|a)$. The output of this DMC forms the state sequence of a channel with input $x \in \mathcal{X}$, output $y \in \mathcal{Y}$ and transition probability $p(y|x, s)$. The encoder $\mathcal{E}_{\text{chan}}$ has non-causal access to the state sequence S^n . $\mathcal{E}_{\text{chan}}$ also needs to convey a private message W_1 . We term this model as *DMADSC with additional private message*. More generally, one can consider common reconstruction (CR) of the state process [8], wherein $\mathcal{E}_{\text{chan}}$ and the decoder must agree on a reconstruction of the state viz. \hat{S}^n . We call this model as *DMADSC with CR*. The inputs have to satisfy an average cost constraint defined by a vector function $\gamma : \mathcal{A} \times \mathcal{X} \rightarrow [0, \infty)^2$, where the cost function for sequences is defined as $\gamma(a^n, x^n) = \frac{1}{n} \sum_{i=1}^n \gamma(a_i, x_i)$. Define a single letter distortion measure $d : \mathcal{S} \times \hat{\mathcal{S}} \rightarrow [0, \infty)$ for state reconstruction, where the distortion between sequences is defined as $d(s^n, \hat{s}^n) = \frac{1}{n} \sum_{i=1}^n d(s_i, \hat{s}_i)$. We assume that the distortion measure is bounded and let $D_{\max} = \max_{s \in \mathcal{S}, \hat{s} \in \hat{\mathcal{S}}} d(s, \hat{s})$. We restrict ourselves to finite alphabets $|\mathcal{A}|, |\mathcal{S}|, |\mathcal{X}|, |\mathcal{Y}|, |\hat{\mathcal{S}}|$.

Definition 1. Let Γ be a vector in \mathbb{R}_+^2 . For the DMADSC with additional private message, an $(n, R_1, R_c, \Gamma, \epsilon)$ scheme consists of two encoder maps:

$$A^n(W_c) : [1 : 2^{nR_c}] \rightarrow \mathcal{A}^n,$$

$$X^n(W_c, W_1, S^n) : [1 : 2^{nR_c}] \times [1 : 2^{nR_1}] \times \mathcal{S}^n \rightarrow \mathcal{X}^n,$$

and a decoding map $\psi(Y^n) : \mathcal{Y}^n \rightarrow [1 : 2^{nR_1}] \times [1 : 2^{nR_c}]$ such that for independent and uniformly distributed choices of (W_c, W_1) , we have

$$\mathbb{P}(\psi(Y^n) \neq (W_1, W_c)) \leq \epsilon, \quad (1)$$

$$\sum_{i=1}^n \mathbb{E}[\gamma_k(A_i, X_i)] \leq n(\Gamma_k + \epsilon), k = 1, 2. \quad (2)$$

(where γ_k and Γ_k represent the k th coordinates of γ and Γ .) For the DMADSC with CR, we additionally define a sender quantization map $\phi : \mathcal{S}^n \times \{1, \dots, 2^{nR_c}\} \times \{1, \dots, 2^{nR_1}\} \rightarrow \hat{\mathcal{S}}^n$ and a decoder reconstruction map $\psi_2(Y^n) : \mathcal{Y}^n \rightarrow \hat{\mathcal{S}}^n$ s.t.

$$\mathbb{P}(\psi_2(Y^n) \neq \phi(S^n, W_c, W_1)) \leq \epsilon, \quad (3)$$

$$\mathbb{E}[d(S^n, \psi_2(Y^n))] \leq n(D + \epsilon). \quad (4)$$

We say that a tuple (R_1, R_c, D, Γ) is achievable if an $(n, R_1, R_c, D, \Gamma, \epsilon)$ coding scheme exists for every $\epsilon > 0$ for sufficiently large n . Let $\mathcal{C}_{\text{CR}}^{\text{dmadsc}}$ be the collection of all achievable (R_1, R_c, D, Γ) tuples. Our main result is stated now.

Theorem 2. For the DMADSC $(p(s|a), p(y|x, s))$ with CR, $\mathcal{C}_{\text{CR}}^{\text{dmadsc}}$ is the closure of the union of all (R_1, R_c, D, Γ) tuples such that

$$R_1 \leq I(U; Y|A) - I(U; S|A), \quad (5)$$

$$R_1 + R_c \leq I(A, U; Y) - I(U; S|A), \quad (6)$$

where the union is over distributions of the form $p(a)p(s|a)p(u|a, s)p(x|u, s)p(y|x, s)$ for which there exists a map $\phi : \mathcal{U} \rightarrow \hat{\mathcal{S}}$ such that $\mathbb{E}[d(S, \phi(U))] \leq D$ and $\mathbb{E}[\gamma(A, X)] \leq \Gamma$ (where the cost constraint is to be understood componentwise), with $|\mathcal{U}| \leq |\mathcal{A}| \cdot |\mathcal{X}| \cdot |\mathcal{S}| + 2$. Notice that $U \rightarrow (X, S) \rightarrow Y$ and $A \rightarrow (U, S) \rightarrow X$ are Markov chains.

Proof. The proof of achievability is given in Appendix A, while the proof of converse is given in Appendix B. ■

Remark 3. For the special case of no CR (i.e. $D \geq D_{\max}$) and $R_c = 0$ ($W_c = 0$), the private message capacity is

$$R_1 \leq \max_{a \in \mathcal{A}, p(u|a, s), p(x|u, s)} [I(U; Y|A = a) - I(U; S|A = a)],$$

which corresponds to selecting the action sequence that leads to the maximum Gelfand-Pinsker rate. On the other hand, when $D \geq D_{\max}$ and $R_1 = 0$ ($W_1 = 0$), the common message capacity is

$$R_c \leq \max_{p(a), p(u|a, s), p(x|u, s)} [I(A, U; Y) - I(U; S|A)],$$

which is nothing but the characterization in [3, Theorem 1].

Remark 4. For the special case of no private message, i.e. $W_1 = 0$, Theorem 2 recovers the common message (DMC) case with CR in [7, Proposition 2].

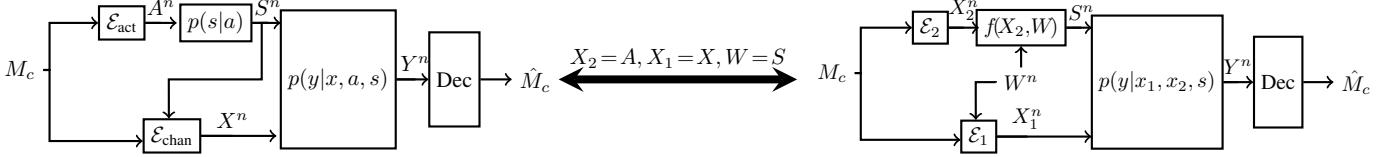


Fig. 2. Relation to the setting of [4], [5]

III. DEGRADED MESSAGE SETS MAC WITH CR

In this section, we show that in the absence of any state reconstruction constraints, i.e. $D \geq D_{max}$, our setting reverts back to the MAC model studied in [4]. As a consequence, we have the following result.

Lemma 5. *The region $\mathcal{C}_{CR}^{dmadsc}$ from Theorem 2 also characterizes the capacity region of the MAC setting in [4], [5] with additional common reconstruction constraints.*

Proof. For simplicity, we show the equivalence when $W_1 = 0$, i.e. no private message- see Fig. 2 for an illustration. The capacity for a channel with action dependent states is given as [3]

$$C_{Weiss} = \max[I(U, A; Y) - I(U; S|A)], \quad (7)$$

where the maximization is over pmfs of the form $p(a)p(s|a)p(u, x|a, s)p(y|x, a, s)$. Now consider the common message capacity of a DM-MAC with asymmetric state information [5]

$$C_{Zaidi} = \max[I(U, X_2; Y) - I(U, X_2; S)], \quad (8)$$

where the maximization is over pmfs of the form $p(x_2)p(s)p(u, x_1|x_2, s)p(y|x_1, x_2, s)$. Suppose we view the action encoder to be playing the role of the MAC encoder that does not observe the state, i.e. we rename $A = X_2$. Since $S \perp\!\!\!\perp X_2$, it is immediate that Weissman's setting is an upper bound for Zaidi's setting. In other words,

$$C_{Zaidi} \leq C_{Weiss}. \quad (9)$$

On the other hand, we can show that any achievable rate in Weissman's setting can also be achieved in Zaidi's setting as follows. Since S is produced from A according to a DMC $p(s|a)$, by functional representation lemma [15], the output of the DMC can be written as

$$S = f(A, W), \quad (10)$$

where $f(\cdot)$ is a deterministic function and $A \perp\!\!\!\perp W$. Note that W^n would be an i.i.d. sequence. We now view W^n to be the state which is known to encoder one. Since the message is common to both encoders, encoder one also knows A^n . Hence it can construct $S^n = f(A^n, W^n)$, and thus obtains S^n as well, as in Weissman's setting. Hence we conclude that

$$C_{Zaidi} \geq C_{Weiss}. \quad (11)$$

This in turn implies that

$$C_{Zaidi} = C_{Weiss}. \quad (12)$$

Thus the settings are equivalent. Clearly, this argument holds even in the presence of an additional private message at the

channel encoder, which allows us to characterize the capacity region for CR in the MAC setting of [4], [5]. ■

From the established equivalence, we also have the following result regarding feedback from the channel output to the channel encoder.

Remark 6. *The capacity region in Theorem 2 remains unchanged even if the channel output Y^{i-1} is causally fed back to the channel encoder.*

Indeed, this is consistent with the observation in [4] that causal feedback of Y^{i-1} to the informed encoder does not enlarge the capacity region.

IV. GAUSSIAN SETTING

The Gaussian action-dependent state channel [3] (GADSC) is given by

$$Y = X + A + W + Z, \quad (13)$$

with $W \sim \mathcal{N}(0, \sigma_W^2)$ and $Z \sim \mathcal{N}(0, \sigma_Z^2)$, $W \perp\!\!\!\perp Z$, and A and X being constrained in average power to P_A and P_X respectively. The state itself is given by $S = A + W$, with A being independent of W . There are two messages to be conveyed, a common message $M_c \in [1 : 2^{nR_c}]$ and a private message $M_1 \in [1 : 2^{nR_1}]$ at the channel encoder, in addition to CR of the state. We have the following theorem.

Theorem 7. *For the GADSC with additional private message and CR constraints, the capacity region is achieved by appropriate jointly Gaussian choices of $p(a), p(u, x|a, s)$ in Theorem 2. Specifically, the capacity region \mathcal{C}_{CR}^{gadsc} is given by*

$$\mathcal{C}_{CR}^{gadsc} = \bigcup \{(R_1, R_c) | R_1 \leq R_{priv}, R_1 + R_c \leq R_{sum}\}, \quad (14)$$

where the union is over $\rho_1, \rho_2 \in [-1, 1]$ satisfying $\rho_1^2 + \rho_2^2 \leq 1$ and $D \geq \frac{(1-\rho_2^2)P_X\sigma_W^2}{P_X + \alpha^2\sigma_W^2 + 2\alpha\rho_2\sqrt{P_X\sigma_W^2}}$, for the rate functions (R_{priv}, R_{sum}) as in (18), (19) and α as in (22).

Proof. We will prove this theorem using the single-letter expression in Theorem 2. While Theorem 2 was shown for the discrete memoryless case, it does hold for the Gaussian case as well. The achievability of Theorem 2 for the Gaussian case follows from an application of the discretization procedure [15, Sec. 3.4.1]. But note that the converse for the discrete memoryless case assumed finite $|\hat{\mathcal{S}}|$ (see (32), i.e. Fano's inequality for $\hat{\mathcal{S}}^n$). However, without loss of generality, we can restrict attention to an exponential number of agreed reconstructions, i.e. $|\hat{\mathcal{S}}^n| = \mathcal{O}(2^{nc})$, where c is a constant, as explained next. A similar argument was also employed in one of our recent works [16] on CR over a Gaussian broadcast channel.

Remark 8. It can be shown that if there exists a scheme without any cardinality bounds on the reproductions which achieves (R_1, R_c, D) , then for any $\delta > 0$, $(R_1, R_c, D+\delta)$ can be achieved using a scheme where the reproductions are confined to lie in an alphabet of size which scales as $2^{nc(\delta)}$, where $c(\delta)$ is a constant. This can be proven using, for instance, a scalar quantizer whose average distortion is δ and whose rate scales with n as $2^{nc(\delta)}$. Thus the converse of Thm. 2 also applies for the Gaussian case, and we can work with single-letter expressions, as opposed to the multi-letter arguments in [16].

We now prove the rest of the converse starting with expressions (5) and (6) from Section II:

$$\begin{aligned} R_1 &\leq h(Y|A) - h(Y|A, U) - h(S|A) + h(S|A, U) \\ &= h(Y|A) - h(Y|A, U, S) - h(S|A) + h(S|Y, A, U) \\ &\leq h(Y|A) - h(Y|A, U, S, X) - h(S|A) + h(S|Y, A, U) \\ &= h(Y|A) - h(Z) - h(S|A) + h(S|Y, A, U) \\ &\leq \frac{1}{2} \log \left[\left(\frac{\sigma_{Y|A}^2}{\sigma_Z^2} \right) \left(\frac{\sigma_{S|Y,A,U}^2}{\sigma_W^2} \right) \right]. \end{aligned} \quad (15)$$

Similarly, for the sum rate

$$\begin{aligned} R_1 + R_c &\leq h(Y) - h(Y|U, A) - h(S|A) + h(S|A, U) \\ &\leq h(Y) - h(Z) - h(S|A) + h(S|Y, A, U) \\ &\leq \frac{1}{2} \log \left[\left(\frac{\sigma_Y^2}{\sigma_Z^2} \right) \left(\frac{\sigma_{S|Y,A,U}^2}{\sigma_W^2} \right) \right]. \end{aligned} \quad (16)$$

where the variance terms in (15) and (16) are defined as

$$\sigma_{S|Y,A,U}^2 = \min_{\alpha, \beta, \gamma} \mathbb{E}[S - \alpha Y - \beta A - \gamma U]^2. \quad (17)$$

Thus it can be seen that, via the differential entropy maximizing property of Gaussian random variables for a given variance, the optimal auxiliary U must be jointly Gaussian with (A, S, X, Y) . This completes the proof of converse. For the achievability, we choose $p(a), p(u, x|a, s)$ as follows. The action input is chosen as $A \sim \mathcal{N}(0, P_A)$. We take the channel input to be

$$X = \rho_1 \sqrt{\frac{P_X}{P_A}} A + \rho_2 \sqrt{\frac{P_X}{\sigma_W^2}} W + G, \quad (20)$$

where ρ_1 and ρ_2 satisfy $\rho_1^2 + \rho_2^2 \leq 1$, and $G \sim \mathcal{N}(0, (1 - \rho_1^2 - \rho_2^2)P_X)$ is independent of (A, W, Z) . We choose the auxiliary random variable as follows:

$$U = \delta X + A + \alpha \delta W, \quad (21)$$

where $\delta = -1 / \left(\rho_1 \sqrt{\frac{P_X}{P_A}} \right)$ and the coefficient α is chosen as

$$\alpha = \frac{(1 - \rho_1^2 - \rho_2^2)P_X - \rho_2 \sqrt{\frac{P_X}{\sigma_W^2}} \sigma_Z^2}{(1 - \rho_1^2 - \rho_2^2)P_X + \sigma_Z^2}. \quad (22)$$

The common reconstruction constraint is as follows:

$$\begin{aligned} D &\geq \mathbb{E}[S - \hat{S}]^2 = \text{Var}[S|U, A] \\ &= \frac{(1 - \rho_2^2)P_X \sigma_W^2}{P_X + \alpha^2 \sigma_W^2 + 2\alpha \rho_2 \sqrt{P_X \sigma_W^2}}. \end{aligned}$$

Now on evaluating the terms in (15), (16) with the above jointly Gaussian choices, we arrive at the rate constraints in (18), (19). \blacksquare

Remark 9. Note that the connection between the current setup and that of [4] established in Section III implies that Theorem 7 also characterizes the rates versus distortion trade-off for common reconstructions in the Gaussian model of [4], a result which was hitherto unknown.

Remark 10. For the GADSC with only a common message and CR constraints considered in [7], the capacity characterization can be obtained by setting $W_1 = 0$ i.e. $R_1 = 0$ in Theorem 7, which is given by (23). Note that [7] only gave an achievable region, while its optimality was not established therein.

$$C(D) = \max_{\substack{(\rho_1, \rho_2) \\ \rho_1^2 + \rho_2^2 \leq 1, D \geq \frac{(1 - \rho_2^2)P_X \sigma_W^2}{P_X + \alpha^2 \sigma_W^2 + 2\alpha \rho_2 \sqrt{P_X \sigma_W^2}}}} R_{\text{sum}}(\rho_1, \rho_2). \quad (23)$$

In fact, our converse proof here relies upon the argument presented in Remark 8.

Remark 11. On further setting $D \geq \sigma_W^2$, i.e. the case of no CR constraints, the characterization simplifies to

$$C = \max_{\substack{(\rho_1, \rho_2) \\ \rho_1^2 + \rho_2^2 \leq 1}} R_{\text{sum}}(\rho_1, \rho_2), \quad (24)$$

which is a complete characterization of the Gaussian action dependent channel. This can also be inferred from the results in [4, Theorem 7] and [5, Theorem 4].

V. CONCLUSION

We studied a generalization of the action-dependent channel with multiple messages and CR constraints, and derived complete characterizations for both discrete memoryless as well as Gaussian models. As a result, we obtained the capacity for a Gaussian action dependent model with only common message and CR. Furthermore, we proved an equivalence between our setting and degraded message sets MAC models in literature, which in turn settles the capacity region of the corresponding MAC model with CR constraints.

APPENDIX A ACHIEVABILITY PROOF FOR THEOREM 2

Proof of Achievability: The achievability is proven using a combination of Gelfand-Pinsker (GP) coding and superposition coding. Here, an action codebook is built first based on the common message. Then for each a^n (action) sequence, a conditional GP codebook U shall be generated according to the private message.

Codebook Generation: Fix the p.m.f. $p(a)p(u|a, s)p(x|u, s)$. Randomly and independently generate 2^{nR_c} sequences $a^n(w_c)$, $w_c \in [1 : 2^{nR_c}]$ i.i.d. according to $\prod_{i=1}^n p_A(a_i)$. For each $a^n(w_c)$ sequence, randomly and conditionally independently generate $2^{n(R_1+R')}$ sequences $u^n(w_c, w_1, j)$ for $w_1 \in [1 : 2^{nR_1}]$ and $j \in [1 : 2^{nR'}]$, i.i.d. according to $\prod_{i=1}^n p_{U|A}(u_i|a_i(w_c))$. Let $\mathcal{B}_U(w_1)$ be the set of sequences within the bin indexed by $w_1 \in [1 : 2^{nR_1}]$ in the U -codebook.

Encoding: Fix $\epsilon' > 0$. Given $w_c \in [1 : 2^{nR_c}]$, the action encoder selects the corresponding sequence $a^n(w_c)$ in the A codebook. The state sequence S^n is generated in response to the action sequence via the channel $p(s|a)$. Given $w_1 \in$

$$R_{priv}(\rho_1, \rho_2) = \frac{1}{2} \log \left[\left(\frac{\sigma_Z^2 + (1 - \rho_1^2 - \rho_2^2)P_X}{\sigma_Z^2} \right) \right], \quad (18)$$

$$R_{sum}(\rho_1, \rho_2) = \frac{1}{2} \log \left[\left(\frac{\sigma_Z^2 + (1 - \rho_1^2 - \rho_2^2)P_X}{\sigma_Z^2} \right) \left(\frac{P_X + P_A + \sigma_Z^2 + \sigma_W^2 + 2\rho_1\sqrt{P_X P_A} + 2\rho_2\sqrt{P_X \sigma_W^2}}{(1 - \rho_1^2)P_X + \sigma_Z^2 + \sigma_W^2 + 2\rho_2\sqrt{P_X \sigma_W^2}} \right) \right]. \quad (19)$$

$[1 : 2^{nR_1}]$, the channel encoder picks the least index j such that $(a^n(w_c), u^n(w_c, w_1, j), s^n) \in \mathcal{T}_{\epsilon'}^n(U, A, S)$. An error is declared if no such index is found. The channel encoder then draws x^n i.i.d. conditionally given (u^n, s^n) according to $\prod_{i=1}^n p(x_i|u_i, s_i)$, and sends it. The channel encoder also generates its state reconstruction $\hat{S}_i = \phi(u_i)$, $i \in [1 : n]$.

Decoding: Let $\epsilon > \epsilon'$. We use simultaneous decoding. The decoder declares that (\hat{w}_c, \hat{w}_1) is sent if it is the unique message pair such that $(a^n(\hat{w}_c), u^n(\hat{w}_c, \hat{w}_1, \hat{j}), y^n) \in \mathcal{T}_{\epsilon}^n(U, A, Y)$ for some $\hat{j} \in \mathcal{B}_U(\hat{w}_1)$. An error is declared otherwise.

Error Analysis: Assume without loss of generality that the messages $W_1 = 1$ and $W_c = 1$ were sent, and the index of the chosen U^n sequence is J . The encoding error events are

$$\mathcal{E}_1 = \{(A^n(1), U^n(1, 1, j), S^n) \notin \mathcal{T}_{\epsilon'}^n \forall j \in [1 : 2^{nR'}]\}. \quad (25)$$

The decoding error events at the receiver can be enumerated as

$$\mathcal{E}_2 = \{(A^n(1), U^n(1, w_1, j), Y^n) \in \mathcal{T}_{\epsilon}^n$$

for some $w_1 \neq 1, j \in \mathcal{B}_U(w_1)\}$,

$$\mathcal{E}_3 = \{(A^n(w_c), U^n(w_c, w_1, j), Y^n) \in \mathcal{T}_{\epsilon}^n(U, A, Y)$$

for some $w_c \neq 1, w_1 \neq 1, j \in \mathcal{B}_U(w_1)\}$,

$$\mathcal{E}_4 = \{(A^n(w_c), U^n(w_c, 1, J), Y^n) \in \mathcal{T}_{\epsilon}^n \text{ for some } w_c \neq 1\}.$$

By the covering lemma [15], we have,

$$\mathbb{P}(\mathcal{E}_1) \xrightarrow{n \rightarrow \infty} 0 \text{ if } R' \geq I(U; S|A) + \delta(\epsilon'). \quad (26)$$

By the packing lemma [15], we can write

$$\mathbb{P}(\mathcal{E}_2) \xrightarrow{n \rightarrow \infty} 0 \text{ if } R_1 + R' \leq I(U; Y|A) - \delta(\epsilon), \quad (27)$$

$$\mathbb{P}(\mathcal{E}_3), \mathbb{P}(\mathcal{E}_4) \xrightarrow{n \rightarrow \infty} 0 \text{ if } R_c + R_1 + R' \leq I(U, A; Y) - \delta(\epsilon). \quad (28)$$

On eliminating R' from equations (26)–(28) via Fourier-Motzkin elimination, we arrive at the rate constraints:

$$R_1 \leq I(U; Y|A) - I(U; S|A) - \delta(\epsilon') - \delta(\epsilon), \quad (29)$$

$$R_1 + R_c \leq I(U, A; Y) - I(U; S|A) - \delta(\epsilon') - \delta(\epsilon). \quad (30)$$

Input Cost Analysis: From the encoding error analysis,

$$\begin{aligned} \mathbb{P}(\mathcal{E}_1) &= \mathbb{P}((U^n(W_1, W_c, J), S^n, A^n(W_c)) \notin \mathcal{T}_{\epsilon'}^n) \\ &= \mathbb{P}((U^n(W_1, W_c, J), S^n, X^n, A^n(W_c)) \notin \mathcal{T}_{\epsilon'}^n) \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

But if $(x^n, a^n) \in \mathcal{T}_{\epsilon'}^n$, then we have $\gamma(x^n, a^n) \leq \Gamma$ by the typical average lemma [15]. Hence we have

$$\begin{aligned} \mathbb{E}[\gamma(X^n, A^n)] &= \mathbb{P}(\mathcal{E}_1) \mathbb{E}[\gamma(X^n, A^n)|\mathcal{E}_1] + \mathbb{P}(\mathcal{E}_1^c) \mathbb{E}[\gamma(X^n, A^n)|\mathcal{E}_1^c] \\ &\leq \mathbb{P}(\mathcal{E}_1) \Gamma_{max} + \mathbb{P}(\mathcal{E}_1^c) \Gamma, \end{aligned}$$

where $\Gamma_{max} = \max_{a \in \mathcal{A}, x \in \mathcal{X}} \gamma(a, x)$

$$\implies \limsup_{n \rightarrow \infty} \mathbb{E}[\gamma(X^n, A^n)] \leq \Gamma, \text{ since } \mathbb{P}(\mathcal{E}_1^c) \xrightarrow{n \rightarrow \infty} 1.$$

Distortion Analysis: The mentioned distortion can be obtained by making the estimate on a per-letter basis. Since $\phi(U)$ satisfies the distortion constraint, it follows from the random codebook construction that as $n \rightarrow \infty$, we have

$$\mathbb{E}[d(S^n, \phi(U^n))] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[d(S_i, \phi(U_i))]$$

$$\xrightarrow{n \rightarrow \infty} \mathbb{E}[d(S, \phi(U))] \leq D. \quad (31)$$

Also, since both the encoder as well as the decoder (assuming correct decoding) can generate $\hat{S}_i = \phi(U_i)$, $i \in [1 : n]$, the common reconstruction constraint is met.

APPENDIX B CONVERSE PROOF FOR THEOREM 2

Proof of Converse: By Fano's inequality, we can write:

$$H(W_1, W_c, \hat{S}^n | Y^n) \leq n\epsilon_n, \text{ where } \epsilon_n \xrightarrow{n \rightarrow \infty} 0. \quad (32)$$

We suppress the $n\epsilon_n$ terms in the sequel. We can write the following chain of inequalities for the private rate:

$$\begin{aligned} nR_1 &= H(W_1) \stackrel{(a)}{=} H(W_1|W_c) \\ &= H(\hat{S}^n, W_1|W_c) - H(\hat{S}^n|W_1, W_c) \\ &\stackrel{(b)}{=} H(\hat{S}^n, W_1|W_c) - H(\hat{S}^n|W_1, W_c) + H(\hat{S}^n|W_1, W_c, S^n) \\ &\stackrel{(c)}{\approx} I(\hat{S}^n, W_1; Y^n|W_c, A^n) - I(\hat{S}^n; S^n|W_1, W_c, A^n) \\ &\stackrel{(d)}{=} I(\hat{S}^n, W_1; Y^n|W_c, A^n) - I(\hat{S}^n, W_1; S^n|W_c, A^n) \\ &= \sum_{i=1}^n \left[I(\hat{S}^n, W_1; Y_i|W_c, Y^{i-1}, A^n) \right. \\ &\quad \left. - I(\hat{S}^n, W_1; S_i|W_c, A^n, S_{i+1}^n) \right] \\ &\stackrel{(e)}{\leq} \sum_{i=1}^n \left[I(\hat{S}^n, W_1, W_c, Y^{i-1}, A^{i-1}, A_{i+1}^n; Y_i|A_i) \right. \\ &\quad \left. - I(\hat{S}^n, W_1, W_c, S_{i+1}^n, A^{i-1}, A_{i+1}^n; S_i|A_i) \right] \\ &\stackrel{(f)}{=} \sum_{i=1}^n \left[I(\hat{S}^n, W_1, W_c, Y^{i-1}, S_{i+1}^n, A^n; Y_i|A_i) \right. \\ &\quad \left. - I(\hat{S}^n, W_1, W_c, Y^{i-1}, S_{i+1}^n, A^n; S_i|A_i) \right] \\ &\stackrel{(g)}{=} \sum_{i=1}^n [I(U_i; Y_i|A_i) - I(U_i; S_i|A_i)], \end{aligned} \quad (33)$$

where (a) follows since $W_1 \perp\!\!\!\perp W_c$, (b) follows since \hat{S}^n is a deterministic function of (W_1, W_c, S^n) , (c) follows from Fano's inequality and since W_c determines A^n , (d) follows since $(W_1, W_c) \rightarrow A^n \rightarrow S^n$, (e) follows since $S_i \rightarrow A_i \rightarrow (W_c, S_{i+1}^n, A^{i-1}, A_{i+1}^n)$, (f) follows from the Csiszár sum lemma and (g) follows with a choice of $U_i = (\hat{S}^n, W_1, W_c, Y^{i-1}, S_{i+1}^n, A^n)$ that satisfies the Markov conditions $U_i \rightarrow (X_i, S_i) \rightarrow Y_i$ and $A_i \rightarrow (U_i, S_i) \rightarrow X_i$. Now by introducing a time-sharing random variable Q , independent of everything else and uniformly drawn on $[1 : n]$, we have

$$\begin{aligned} nR_1 &\leq n[I(U_Q; Y_Q|A_Q, Q) - I(U_Q; S_Q|A_Q, Q)] \\ &= n[I(U_Q; Y_Q|A_Q, Q) - I(U_Q, Q; S_Q|A_Q)] \end{aligned}$$

$$\leq n [I(U_Q, Q; Y_Q | A_Q) - I(U_Q, Q; S_Q | A_Q)]. \quad (34)$$

Now the proof is completed by replacing $(Q, U_Q) = U, Y_Q = Y, S_Q = S, X_Q = X$ and $A_Q = A$ and noting that the Markov conditions $U \rightarrow (X, S) \rightarrow Y$ and $A \rightarrow (U, S) \rightarrow X$ hold. For the sum rate, consider the following chain of inequalities:

$$\begin{aligned} n(R_1 + R_c) &= H(W_1, W_c) \\ &= H(\hat{S}^n, W_1, W_c) - H(\hat{S}^n | W_1, W_c) \\ &\stackrel{(a)}{=} H(\hat{S}^n, W_1, W_c, A^n) - I(\hat{S}^n; S^n | W_1, W_c, A^n) \\ &\stackrel{(b)}{\approx} I(\hat{S}^n, W_1, W_c, A^n; Y^n) - I(\hat{S}^n; S^n | W_1, W_c, A^n) \\ &\stackrel{(c)}{=} I(\hat{S}^n, W_1, W_c, A^n; Y^n) - I(\hat{S}^n, W_1, W_c; S^n | A^n) \\ &\stackrel{(d)}{\leq} \sum_{i=1}^n \left[I(\hat{S}^n, W_1, W_c, Y^{i-1}, A^n; Y_i) \right. \\ &\quad \left. - I(\hat{S}^n, W_1, W_c, S_{i+1}^n, A^{i-1}, A_{i+1}^n; S_i | A_i) \right] \\ &\stackrel{(e)}{=} \sum_{i=1}^n \left[I(\hat{S}^n, W_1, W_c, Y^{i-1}, S_{i+1}^n, A^n; Y_i) \right. \\ &\quad \left. - I(\hat{S}^n, W_1, W_c, Y^{i-1}, S_{i+1}^n, A^n; S_i | A_i) \right] \\ &\stackrel{(f)}{=} \sum_{i=1}^n [I(U_i, A_i; Y_i) - I(U_i; S_i | A_i)], \end{aligned} \quad (35)$$

where (a) follows since \hat{S}^n is a deterministic function of (W_1, W_c, S^n) and since W_c determines A^n , (b) follows from Fano's inequality, (c) follows since $(W_1, W_c) \rightarrow A^n \rightarrow S^n$, (d) follows since $S_i \rightarrow A_i \rightarrow (S_{i+1}^n, A^{i-1}, A_{i+1}^n)$, (e) follows from the Csiszár sum lemma and (f) follows since $U_i = (\hat{S}^n, W_1, W_c, Y^{i-1}, S_{i+1}^n, A^n)$. Now by introducing a time-sharing RV Q , as in (34), we have

$$n(R_1 + R_c) \leq n [I(U_Q, Q, A_Q; Y_Q) - I(U_Q, Q; S_Q | A_Q)].$$

For the bound on input costs, we proceed as follows:

$$\begin{aligned} \mathbb{E}[\gamma_k(A, X)] &= \sum_{a,x} \mathbb{P}(A = a, X = x) \gamma_k(a, x) \\ &= \sum_{a,x} \mathbb{P}(A_Q = a, X_Q = x) \gamma_k(a, x) \\ &= \sum_{a,x} \sum_{i=1}^n \mathbb{P}(A_Q = a, X_Q = x | Q = i) \mathbb{P}(Q = i) \gamma_k(a, x) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{a,x} \mathbb{P}(A_i = a, X_i = x) \gamma_k(a, x) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\gamma_k(A_i, X_i)] \leq (\Gamma_k + \epsilon), k = 1, 2, \end{aligned} \quad (36)$$

where the last step follows from the fact that (2) holds for any successful $(n, R_1, R_c, D, \Gamma, \epsilon)$ scheme. We next consider the bound on the distortion. Let \hat{S}_d^n be the receiver's reconstruction of the state sequence. Define the event $\mathcal{B} := \{\hat{S}^n = \hat{S}_d^n\}$. Notice that $P(\mathcal{B}) \geq 1 - \epsilon, \forall \epsilon > 0$. Hence

$$n(D + \epsilon) \geq \sum_{i=1}^n \mathbb{E}[d(S_i, \hat{S}_{di})] = \mathbb{E}_{\mathcal{B}} \left[\sum_{i=1}^n \mathbb{E}[d(S_i, \hat{S}_{di})] \middle| \mathcal{B} \right]$$

$$\begin{aligned} &\geq P(\mathcal{B}) \sum_{i=1}^n \mathbb{E}[d(S_i, \hat{S}_i)] \stackrel{(a)}{\geq} (1 - \epsilon) \sum_{i=1}^n \mathbb{E}[d(S_i, \phi_i(U_i))] \\ &= n(1 - \epsilon) \mathbb{E}_Q [\mathbb{E}[d(S_Q, \phi_Q(U_Q))] | Q] \\ &= n(1 - \epsilon) \mathbb{E}[d(S_Q, \phi_Q(U_Q))] = n(1 - \epsilon) \mathbb{E}[d(S, \phi_Q(U_Q))] \\ &\stackrel{(b)}{=} n(1 - \epsilon) \mathbb{E}[d(S, \phi(U_Q, Q))] \stackrel{(c)}{=} n(1 - \epsilon) \mathbb{E}[d(S, \phi(U))], \end{aligned}$$

where (a) follows by taking $\phi_i(U_i) = \hat{S}_i$, (b) follows by defining $\phi : (Q, U_Q) \mapsto \phi_Q(U_Q)$ and (c) follows since $(U_Q, Q) = U$. Finally, the bound on $|\mathcal{U}|$ follows by an application of the support lemma [15, Appendix C].

ACKNOWLEDGEMENTS

The work was supported in part by the Bharti Centre for Communication, IIT Bombay, the grant 17ISROCO08 from the ISRO-IITB Space Technology Cell and the grant EMR/2016/005847 from the DST, India.

REFERENCES

- [1] C. E. Shannon, "Channels with side information at the transmitter," *IBM journal of Research and Development*, vol. 2, no. 4, pp. 289–293, 1958.
- [2] S. Gelfand and M. Pinsker, "Coding for channels with random parameters," *Probl. Contr. Inform. Theory*, vol. 9, no. 1, pp. 19–31, 1980.
- [3] T. Weissman, "Capacity of channels with action-dependent states," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5396–5411, 2010.
- [4] A. Somekh-Baruch, S. Shamai, and S. Verdú, "Cooperative multiple-access encoding with states available at one transmitter," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4448–4469, 2008.
- [5] A. Zaidi, P. Piantanida, and S. Shamai, "Capacity region of cooperative multiple-access channel with states," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6153–6174, 2013.
- [6] ———, "Wyner-Ziv type versus noisy network coding for a state-dependent mac," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 1682–1686.
- [7] B. Ahmadi and O. Simeone, "On channels with action-dependent states," in *Information Theory Workshop (ITW), 2012 IEEE*, 2012, pp. 167–171.
- [8] Y. Steinberg, "Coding and common reconstruction," *Information Theory, IEEE Transactions on*, vol. 55, no. 11, pp. 4995–5010, 2009.
- [9] H. Permuter and T. Weissman, "Source coding with a side information vending machine," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4530–4544, 2011.
- [10] H. Asnani, H. Permuter, and T. Weissman, "Probing capacity," *IEEE Transactions on Information Theory*, vol. 57, no. 11, pp. 7317–7332, 2011.
- [11] C. Choudhuri and U. Mitra, "Action dependent strictly causal state communication," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 3058–3062.
- [12] K. Kittichokechai, T. J. Oechtering, and M. Skoglund, "Coding with action-dependent side information and additional reconstruction requirements," *IEEE Transactions on Information Theory*, vol. 61, no. 11, pp. 6355–6367, 2015.
- [13] Y. Steinberg and T. Weissman, "The degraded broadcast channel with action-dependent states," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 596–600.
- [14] Y. Steinberg, "The degraded broadcast channel with non-causal action-dependent side information," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2965–2969.
- [15] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.
- [16] V. Ramachandran, S. R. B. Pillai, and V. M. Prabhakaran, "State-Dependent Gaussian Broadcast Channel with Common State Reconstructions," in *Proceedings of International Symposium on Information Theory and its Applications (ISITA), Singapore*, 2018, pp. 690–694.

Conditions for Optimality of Superposition Coding in Discrete Memoryless Broadcast Channels

Harikumar Krishnamurthy, Parikshit Hegde and Andrew Thangaraj

Department of Electrical Engineering

Indian Institute of Technology Madras

Chennai, India

ee14b129,ee14b123, andrew@ee.iitm.ac.in

Abstract—The capacity region of general discrete-memoryless broadcast channels (DMBCs) with two receivers is an open problem of considerable research interest. The optimality of superposition coding in three specific cases of the DMBC is considered. For a DMBC with binary input, symmetric output and output cardinality at most 3, superposition coding is shown to be optimal. For equal-capacity DMBCs with any input cardinality, superposition coding is shown to be suboptimal if each channel has a capacity-achieving input distribution that is not capacity-achieving for the other channel. For an equal-capacity DMBC with binary input, superposition coding is shown to be optimal if and only if the two channels are *more-capable comparable* even without output symmetry. These results improve upon the previously known conditions for optimality of superposition coding in DMBCs.

Index Terms—Broadcast channels, Capacity region, Superposition coding

I. INTRODUCTION

A. Preliminaries and notation

Consider a two-receiver, discrete memoryless broadcast channel [1] [2] with input $X \in \mathcal{X}$ resulting in outputs $Y_i \in \mathcal{Y}_i$ to Receiver i for $i = 1, 2$. The channel to Receiver i , denoted $X \rightarrow Y_i$, is defined through the transition probabilities

$$P(Y_i|X) \triangleq [\Pr(Y_i=y|X=x)]_{x \in \mathcal{X}, y \in \mathcal{Y}_i}.$$

A $(2^{nR_1}, 2^{nR_2}, n)$ code consists of an encoder that maps messages $M_i \in [2^{nR_i}] \triangleq \{1, 2, \dots, 2^{nR_i}\}$, $i = 1, 2$, to a codeword in \mathcal{X}^n and two decoders that map received values in \mathcal{Y}_i^n to $\hat{M}_i \in [2^{nR_i}]$. Assuming that M_i are uniformly and independently chosen, the probability of error is defined as

$$P_e^{(n)} = \Pr(\{M_1 \neq \hat{M}_1\} \cup \{M_2 \neq \hat{M}_2\}).$$

A rate pair (R_1, R_2) is achievable if a sequence of $(2^{nR_1}, 2^{nR_2}, n)$ codes exist with $P_e^{(n)} \rightarrow 0$ as $n \rightarrow \infty$. The closure of the set of all achievable rate pairs forms the capacity region. Characterising the capacity region in single letter (independent of n) is an open problem for general discrete-memoryless broadcast channels, and inner/outer bounds have been the topic of several recent papers [3]–[10].

Entropy and mutual information will be in the units of nats. The capacity of the channel $X \rightarrow Y_i$, denoted C_i , is defined as $C_i \triangleq \max_{P(X)} I(X; Y_i)$. For binary-input channels

$(X \in \{0, 1\})$, we use the notation $I_i(x) \triangleq I(X; Y_i)|_{P(X=0)=x}$ for $x \in [0, 1]$. Further, $I_d(x) \triangleq I_1(x) - I_2(x)$. A binary-input channel $X \rightarrow Y$ with $Y \in \mathcal{Y}$ is said to have *symmetric output* if $\mathcal{Y} = \{0, \pm 1, \dots, \pm(m-1), \pm m\}$ for a positive integer m and $P(k|0) = P(-k|1)$ for $k \in \{-m, \dots, m\}$. We will consider equal-capacity broadcast channels satisfying $C_1 = C_2$, binary-input broadcast channels ($X \in \{0, 1\}$) and binary-input symmetric output (BISO) broadcast channels ($X \rightarrow Y_i$ is BISO for $i = 1, 2$) especially in this work.

B. Partial ordering of channels

We will use two partial orders on channels.

(1) A channel $X \rightarrow Y_1$ is said to be *more capable* than a channel $X \rightarrow Y_2$ if $I(X; Y_1) \geq I(X; Y_2)$ for all input distributions $P(X)$. [11]

(2) A BISO channel $X \rightarrow Y_1$ is said to be *e-less-noisy* than a BISO channel $X \rightarrow Y_2$ if $I_d(x) \leq C_1 - C_2$ for all $x \in [0, 1]$.

Note that (2) above is a sufficient condition for *essentially-less-noisy* ordering introduced in [4] for the case of symmetric channels. We require only this version of essentially less noisy comparison for our results.

C. Achievable regions

Three achievable regions are of interest in this work.

(1) The *time-division* (TD) region is given by

$$R_1/C_1 + R_2/C_2 \leq 1. \quad (1)$$

(2) The *superposition* region is defined as the convex hull of the following regions:

$$R_1 \leq I(U; Y_1), R_2 \leq I(X; Y_2|U), \quad (2)$$

$$R_1 + R_2 \leq I(X; Y_2), \quad (3)$$

where U is an auxiliary random variable and $U - X - (Y_1, Y_2)$ is a Markov chain. If $C_1 = C_2$, the superposition region coincides with the TD region. The superposition region is optimal if the two channels are either more-capable or e-less-noisy comparable [12] [4].

(3) The *randomized time-division* (RTD) region is defined as the convex hull of the following regions:

$$R_1 \leq I(W; Y_1) + P(W=0)I(X; Y_1|W=0), \quad (4)$$

$$R_2 \leq I(W; Y_2) + P(W=1)I(X; Y_2|W=1), \quad (5)$$

$$\begin{aligned} R_1 + R_2 &\leq \min\{I(W; Y_1), I(W; Y_2)\} + \\ &P(W=0)I(X; Y_1|W=0) + P(W=1)I(X; Y_2|W=1), \end{aligned} \quad (6)$$

where W is a binary random variable and $W - X - (Y_1, Y_2)$ is a Markov chain. A strategy to prove suboptimality of superposition coding is to show that the superposition region is contained strictly inside the RTD region.

D. Summary of results

The following are the results in this work.

- (1) Any two BISO channels with 3 outputs or lesser are either more capable comparable or e-less-noisy comparable. Hence, superposition region is the capacity region for BISO broadcast channels with at most 3 outputs for each receiver. This extends [6, Corollary 3] to the case of unequal capacity.
- (2) In a binary-input, equal-capacity broadcast channel, either the two channels are more-capable comparable or superposition coding is not optimal. This extends [6, Corollary 5] to the case when the outputs are not symmetric.

The proof of (1) above is through a lemma limiting the number of roots of $I_d''(x)$ in $[0, 1/2]$ to 1. The proof of (2) is through the following result, which is of independent interest.

(2a) In an equal-capacity broadcast channel, if a capacity-achieving input distribution of one channel is not capacity-achieving in the other channel (and vice versa), then superposition coding is not optimal.

A recent paper on optimality of superposition region is [13]. To the best of our reading, our results here are more direct and specific, and they do not appear to be contained in the results presented in [13], which is concerned with more general conditions involving convex envelopes.

II. BISO BROADCAST CHANNEL WITH 3 OUTPUTS

We will consider a 2-receiver, 3-output BISO broadcast channel(as defined in the Introduction and [6]) with the channels to the two receivers (Channel 1 and 2) being

$$P(Y_1|X) = \begin{bmatrix} 1 - p_1 - e_1 & e_1 & p_1 \\ p_1 & e_1 & 1 - p_1 - e_1 \end{bmatrix}, \quad (7)$$

$$P(Y_2|X) = \begin{bmatrix} 1 - p_2 - e_2 & e_2 & p_2 \\ p_2 & e_2 & 1 - p_2 - e_2 \end{bmatrix}, \quad (8)$$

where $e_i, p_i \geq 0$ and $e_i + p_i \leq 1$ for $i = 1, 2$.

To reduce clutter, we will use the notation $I_d'(0) \triangleq \lim_{x \rightarrow 0+} I_d'(x)$ and likewise for higher-order derivatives as well.

A. Roots of $I_d''(x)$

Denoting $x = P(X = 0)$, the mutual information between X and Y_i can be simplified to the following [1, Exercise 7.13]:

$$I_i(x) = (1 - e_i) h\left(\frac{x(1 - e_i) + p_i(1 - 2x)}{1 - e_i}\right) + c_i, \quad (9)$$

where $h(x) = -x \ln(x) - (1 - x) \ln(1 - x)$ is the binary entropy function and c_i is independent of x . Note that $I_i(0) = I_i(1) = 0$, and $C_i = I_i(1/2)$. Recall the notation $I_d(x) = I_1(x) - I_2(x)$.

The double derivative of $h(x)$ is given by $-\frac{1}{x(1-x)}$. Hence, the double derivative of $I_i(x)$ can be simplified to the following:

$$\begin{aligned} I_i''(x) &= \frac{(1 - e_i - 2p_i)^2}{x(1 - e_i) + p_i(1 - 2x)} \\ &\times \frac{(1 - e_i)}{(1 - x)(1 - e_i) + p_i(1 - 2(1 - x))}. \end{aligned} \quad (10)$$

Using the above, we obtain an expression for $I_d''(x) = I_1''(x) - I_2''(x)$, which is useful to prove the following lemma.

Lemma 1. $I_d''(x)$ has either at most one zero in $[0, 1/2]$ or $I_d(x) = 0$ for all x .

Proof. From (10) it is clear that $I_d''(x)$ is of the following form:

$$I_d''(x) = \frac{N(x)}{D(x)}, \quad (11)$$

where $N(x)$ is quadratic and satisfies $N(x) = N(1 - x)$. Therefore, if a is a root of $N(x)$, then $1 - a$ is also a root of $N(x)$.

Therefore, if there exists more than one root for $N(x)$ in $[0, 1/2]$, there exists more than one root in $(1/2, 1]$. Hence $N(x)$ will have more than two roots, which is a contradiction as $N(x)$ is quadratic in x unless, $I_d''(x) = 0, \forall x$. If $I_d''(x) = 0 \forall x$, $I_d(x)$ is linear. But $I_d(0) = I_d(1) = 0$ and hence $I_d(x) = 0 \forall x$.

Hence $I_d''(x)$ can have at most one zero in $[0, 1/2]$ unless $I_d(x) = 0 \forall x$. \square

Remark 1. If $x = 1/2$ is a root of $N(x)$, then we have a double root at $1/2$, and $I_d(x)$ is either concave or convex in $[0, 1]$.

B. More-capable and e-less-noisy comparison

We will use Lemma 1 to determine conditions under which Channels 1 and 2 are either more-capable comparable or e-less-noisy comparable.

We begin with a sufficient condition for Channel 1 being more capable than Channel 2.

Lemma 2. If $C_1 \geq C_2$ and $I_1'(0) \geq I_2'(0)$, then Channel 1 is more capable than Channel 2.

Proof. We prove this by repeated application of the Intermediate Value Theorem and Rolle's Theorem from basic calculus and the details are given in the Appendix. \square

In fact, the above sufficient condition is necessary as well. This is shown in the following lemma.

Lemma 3. *Channel 1 is more capable than Channel 2 if and only if $C_1 \geq C_2$ and $I'_1(0) \geq I'_2(0)$.*

Proof. The “if” direction is clear from Lemma 2.

When Channel 1 is more capable than Channel 2, by definition $I_1(x) \geq I_2(x)$ or $I_d(x) \geq 0$ for all $x \in [0, 1]$. So,

$$C_1 = I_1(1/2) \geq I_2(1/2) = C_2.$$

Further, since $I_d(0) = 0$, if $I'_1(0) < I'_2(0)$, there exists $0 < a_1$ such that $I_d(a_1) < 0$, which is a contradiction. Hence $I'_1(0) \geq I'_2(0)$. \square

Remark 2. (*Condition for less noisy comparison*)

Another important partial order is the less noisy comparison. Channel 1 is less noisy than Channel 2 if $I_1(x) - I_2(x)$ is concave in $[0, 1]$ [14]. Other than the trivial case when $I_d(x) = 0$ for all x , we can show that Channel 1 is less noisy than Channel 2 iff $C_1 > C_2$ and $I''_d(x)$ does not have any zeros in $(0, 1/2)$.

Next, we consider e-less-noisy comparison, for which the following is a sufficient condition.

Lemma 4. *If $C_1 \geq C_2$ and $I'_1(0) < I'_2(0)$, then Channel 1 is e-less-noisy than Channel 2.*

Proof. A proof is given in the Appendix. \square

C. Result for 3-output BISO broadcast channels

Using Lemmas 3 and 4, we have the following theorem.

Theorem 5. *The two channels in a 3-output BISO broadcast channel are either more-capable comparable or e-less-noisy comparable.*

Proof. When $C_1 \geq C_2$, Channel 1 is either more capable than Channel 2 (when $I'_d(0) \geq 0$) or Channel 1 is e-less noisy than Channel 2 (when $I'_d(0) < 0$) by Lemmas 3 and 4. Similarly, when $C_1 \leq C_2$, either Channel 2 is more capable than Channel 1 or Channel 2 is e-less noisy than Channel 1. \square

Since superposition coding is optimal if the two channels are more-capable or e-less noisy comparable [12] [4], the following corollary is immediate.

Corollary 6. *For a 3-output BISO broadcast channel, superposition coding is optimal.*

D. 4-output BISO broadcast channels

We attempted to extend the same analysis to 4-output BISO broadcast channels. In this case, we obtain a biquadratic polynomial in the numerator of the double derivative $I''_d(x)$. We observed numerically that for most choice of channel parameters, $I''_d(x)$ has only one zero in $(0, 1/2)$. This means that for 4-output BISO broadcast channels, superposition coding is almost always optimal. However, no analytical results have been obtained so far.

III. EQUAL-CAPACITY BROADCAST CHANNEL

We will now consider equal-capacity discrete memoryless broadcast channels $X \rightarrow (Y_1, Y_2)$ where both the receiver channels $X \rightarrow Y_1$ and $X \rightarrow Y_2$ have the same capacity $C_1 = C_2 = C$. There can be two different scenarios based on how capacity-achieving input distributions for the two channels are located relative to each other.

A. Exclusive capacity-achieving input distributions

In this scenario, each channel has at least one capacity-achieving input distribution that is not capacity-achieving for the other channel. Under this scenario, we require no further assumptions on the input and output alphabet cardinality or symmetry.

Lemma 7. *In an equal-capacity broadcast channel, if there exist capacity-achieving input distributions for each channel that are not capacity-achieving for the other channel, superposition coding is not optimal.*

Proof. Let the RV $X_1 \sim p_1$ be such that $I(X_1; Y_1) = C$ and $I(X_1; Y_2) < C$. Similarly, let the RV $X_2 \sim p_2$ be such that $I(X_2; Y_2) = C$ and $I(X_2; Y_1) < C$.

We will prove that superposition coding is not optimal by showing that the superposition region, which is actually the time-division region $R_1 + R_2 \leq C$ for equal-capacity broadcast channels, is a proper subset of the randomized time division (RTD) region in (4)-(6).

Construct random variables $W, X \sim p(w, x)$ such that $X|W=0 \sim p_1$ and $X|W=1 \sim p_2$. The marginal distribution of W can be chosen such that $0 < P(W=0) < 1$. For this distribution, (6) in the RTD region simplifies to the following.

$$R_1 + R_2 \leq \min\{I(W; Y_1), I(W; Y_2)\} + C. \quad (12)$$

For the chosen distribution, we will show that $I(W; Y_1) > 0$ through a proof by contradiction. By a similar argument, $I(W; Y_2) > 0$ as well.

Suppose that $I(W; Y_1) = 0$. This implies that the random variables W and Y_1 are independent. Now, consider the following:

$$\begin{aligned} I(X; Y_1|W=1) &= H(Y_1|W=1) - H(Y_1|X, W=1), \\ &\stackrel{(a)}{=} H(Y_1|W=0) - H(Y_1|X, W=0), \\ &= I(X; Y_1|W=0), \\ &= C, \end{aligned} \quad (13)$$

where (a) follows from the fact that $H(Y_1|W=1) = H(Y_1|W=0)$ because W and Y_1 are independent, and the fact that $H(Y_1|X, W=1) = H(Y_1|X, W=0)$ because $W \rightarrow X \rightarrow Y_1$ is a Markov chain.

Since the random variables W, X are such that $I(X; Y_1|W=1) < C$, (13) above is a contradiction. Therefore, $I(W; Y_1) = \epsilon_1 > 0$. Similarly, $I(W; Y_2) = \epsilon_2 > 0$. Let, $\epsilon = \min\{\epsilon_1, \epsilon_2\}$. Using in (12), we have

$$R_1 + R_2 \leq C + \epsilon, \quad \epsilon > 0. \quad (14)$$

Now, we need to show that a part of the boundary of the region specified above, i.e. the line $R_1 + R_2 = C + \epsilon$, exists in the RTD achievable region. That is, we need to check that intersections with regions specified by the (4) and (5) include a part of the boundary in (14). The sum of (4) and (5) results in

$$R_1 + R_2 \leq I(W; Y_1) + I(W; Y_2) + C.$$

Since $I(W; Y_i) \geq \epsilon$, the RHS above satisfies

$$I(W; Y_1) + I(W; Y_2) + C \geq C + 2\epsilon.$$

Therefore, a part of the line $R_1 + R_2 = C + \epsilon$ exists in the RTD region.

Thus, the superposition region given by $R_1 + R_2 \leq C$ is a proper subset of the RTD region. \square

B. Common capacity-achieving input distribution

In this scenario, the two channels share a common capacity-achieving input distribution. No symmetry assumptions are made on the output. However, for our results, we need to assume that the input is binary. Specifically, recalling the notation $I_i(x)$ for the binary input case, we consider a binary-input broadcast channel with $C_1 = C_2 = I_1(x_c) = I_2(x_c)$ for some $x_c \in (0, 1)$.

The interesting case is when the two channels are not more-capable comparable. In that case, we have the following proposition.

Proposition 8. Consider two binary-input channels that are not more-capable comparable. For every $x_0 \in (0, 1)$, there exist $x_1, x_2 \in (0, 1)$ such that (1) $x_0 = \lambda x_1 + (1 - \lambda)x_2$, for some $\lambda \in (0, 1)$, (2) $I_1(x_1) > I_2(x_1)$, and (3) $I_1(x_2) < I_2(x_2)$.

Proof. Since the two channels are not more-capable comparable, there exist $u_1, u_2 \in (0, 1) \setminus \{x_0\}$ (x_0 can be excluded by continuity of $I_i(x)$) such that $I_1(u_1) > I_2(u_1)$ and $I_1(u_2) < I_2(u_2)$. There are multiple cases to consider depending on where u_1 and u_2 lie with respect to x_0 .

If u_1 and u_2 are on either side of x_0 , the choice of x_1 and x_2 can be made readily.

If u_1 and u_2 are both lesser than x_0 , we pick a $u_3 \in (x_0, 1)$ such that $I_1(u_3) \neq I_2(u_3)$. We claim that this is possible always for channels that are not more-capable comparable, and a proof is given a little later. Now, if $I_1(u_3) > I_2(u_3)$, we choose $x_1 = u_3$ and $x_2 = u_2$; else, if $I_1(u_3) < I_2(u_3)$, we choose $x_1 = u_1$ and $x_2 = u_3$. A similar argument can be made if u_1 and u_2 are both greater than x_0 , and the proof of the proposition is complete.

Claim: If $I_1(x) = I_2(x)$ for x in a sub-interval of $(0, 1)$, then $I_1(x) = I_2(x)$ for $x \in [0, 1]$ making the two channels more-capable comparable.

Proof of claim: When the number of outputs are finite, the mutual information $I_i(x)$ consists of a summation of a finite number of terms of the form $A(x) \log A(x)$, where $A(x)$ is an affine function of x . Therefore, the double derivative $I''_d(x)$ is a rational function of x . That is, it only has a finite number

of zeros. So, if $I''_d(x) = 0$ in a sub-interval of $(0, 1)$, then $I'_d(x) = 0$ for its entire domain $x \in (0, 1)$. This is possible only when $I_d(x) = 0$ for $x \in [0, 1]$. \square

Proposition 8 plays a key role in the following lemma.

Lemma 9. Consider an equal-capacity, binary-input broadcast channel, $X \rightarrow (Y_1, Y_2)$, with $P(X = 0) = x_c$ being a common capacity-achieving input distribution for both channels $X \rightarrow Y_1$ and $X \rightarrow Y_2$. For this broadcast channel, either the two channels are more-capable comparable or superposition coding is not optimal.

Proof. The proof is similar to the proof of Theorem 3 in [6] with the use of Proposition 8 as the starting point. First, if the two channels are more-capable comparable, clearly superposition coding is optimal.

Now, consider that the two channels are not more-capable comparable. Use Proposition 8 with $x_0 = x_c$ to obtain x_1, x_2 and λ . Construct binary random variables W, X as follows: $P(X=0|W=0) = x_1$, $P(X=0|W=1) = x_2$, and $P(W=0) = \lambda$. With this choice, the rest of the proof is very similar to [6, Theorem 3], and we will be very brief.

$$\begin{aligned} & I(W; Y_1) + P(W=0)I(X; Y_1|W=0) \\ & \quad + P(W=1)I(X; Y_2|W=1) \\ & \stackrel{(a)}{=} I(X; Y_1) - I(X; Y_1|W) + P(W=0)I(X; Y_1|W=0) \\ & \quad + P(W=1)I(X; Y_2|W=1) \\ & \stackrel{(b)}{=} I(X; Y_1) + P(W=1)(I(X; Y_2|W=1) - I(X; Y_1|W=1)) \\ & \stackrel{(c)}{=} C + \epsilon_1, \quad \epsilon_1 > 0, \end{aligned}$$

where, (a) follows because $W - X - Y_1$ is a Markov Chain, (b) is obtained by expanding the mutual information, and (c) follows by setting $\epsilon_1 = (1 - \lambda)(I_2(x_2) - I_1(x_2))$.

Similarly, for $\epsilon_2 > 0$,

$$\begin{aligned} & I(W; Y_2) + P(W=0)I(X; Y_2|W=0) \\ & \quad + P(W=1)I(X; Y_1|W=1) \\ & = C + \epsilon_2. \end{aligned}$$

Thus, for $\epsilon = \min\{\epsilon_1, \epsilon_2\}$, (6) for the RTD region reduces to

$$R_1 + R_2 \leq C + \epsilon, \quad \epsilon > 0.$$

For a proof that a part of the boundary above will exist in the RTD region, see the latter half of proof of Lemma 7.

Thus, the superposition region $R_1 + R_2 \leq C$ is a strict subset of the RTD region, and superposition coding is not optimal. \square

C. Equal-capacity, binary-input broadcast channels

We can now state the main result for optimality of superposition coding in equal-capacity, binary-input broadcast channels.

Theorem 10. In an equal-capacity, binary-input broadcast channel, either the two channels are more-capable comparable or superposition coding is not optimal.

Proof. The theorem is proved by considering 2 cases - exclusive capacity-achieving input distributions and common capacity-achieving input distributions. While the two cases are not necessarily mutually exclusive, they exhaust all possibilities. By Lemmas 7 and 9, the proof is complete for both the cases above. \square

IV. CONCLUDING REMARKS

While the general problem of characterizing the capacity region of discrete-memoryless broadcast channels is highly challenging, there are several interesting smaller problems in the area that continue to remain open. One such problem is characterizing the optimality of superposition coding.

In this work, we have made some limited progress in the problem of characterizing optimality of superposition coding by showing that superposition coding is optimal for 3-output, binary-input symmetric channels. The method exploits the analytical properties of derivatives of mutual information and appears to be difficult to extend to 4 or higher outputs.

For the case of equal-capacity broadcast channels, we have taken the approach of considering the nature of capacity-achieving input distributions of the two receiver channels. Based on how the capacity-achieving input distributions occur relative to each other, we are able to characterize optimality of superposition in the case of equal-capacity, binary-input broadcast channels with no output symmetry assumptions. The method here is to study the boundary of the randomized time division region in comparison to that of the superposition region. This method appears to be promising for further work and extensions.

APPENDICES

A. Proof of Lemma 2

We will consider two cases: $I'_d(0) > 0$ and $I'_d(0) = 0$.

Case 1: $C_1 \geq C_2$ and $I'_1(0) > I'_2(0)$

We will prove that Channel 1 is more capable than Channel 2 by contradiction. The proof goes by finding a series of points in $(0, 1/2)$ with suitable properties.

- Suppose Channel 1 is not more capable than Channel 2. By symmetry of BISO channel, there exists $a_2 \in (0, 1/2)$ such that $I_1(a_2) < I_2(a_2)$ resulting in $I_d(a_2) < 0$.
- Since $I'_1(0) > I'_2(0)$ i.e. $I'_d(0) > 0$ and $I_d(0) = 0$, there exists $a_1 \in (0, a_2)$ such that $I_d(a_1) > 0$.
- Since $I_i(x)$ is continuous, by intermediate value theorem, there exists $b_1 \in (a_1, a_2)$ such that $I_d(b_1) = 0$. Since $C_1 \geq C_2$ implies $I_d(1/2) \geq 0$, there exists $b_2 \in (a_2, 1/2]$ such that $I_d(b_2) = 0$.
- Since $I_d(0) = I_d(b_1) = I_d(b_2) = 0$, by Rolle's theorem, there exists c_1, c_2 such that $0 < c_1 < b_1 < c_2 < b_2 \leq 1/2$ and $I'_d(c_1) = 0$ and $I'_d(c_2) = 0$.
- Since $I'_d(1/2) = 0$, by Rolle's theorem on $I'_d(x)$, there exist d_1, d_2 such that $0 < c_1 < d_1 < c_2 < d_2 < 1/2$ and $I''_d(d_1) = 0$ and $I''_d(d_2) = 0$.

This contradicts Lemma 1.

Case 2: If $C_1 \geq C_2$ and $I'_1(0) = I'_2(0)$.

The series of arguments is similar to that of Case 1, and we skip the details.

B. Proof of Lemma 4

We will prove this by contradiction.

- Assume that Channel 1 is not e-less noisy than Channel 2. Then, from the definition of e-less noisy ordering, there exists $a_2 \in (0, 1/2)$ such that $I_d(a_2) > I_d(1/2) > 0$.
- Since $I'_d(0) < 0$, there exists $a_1 \in (0, a_2)$ such that $I_d(a_1) < 0$.
- By intermediate value theorem, there exists $b_1 \in (a_1, a_2)$ such that $I_d(b_1) = 0$. Similarly, there exists $b_2 \in [b_1, a_2)$ such that $I_d(b_2) = I_d(1/2)$.
- Since $I_d(0) = I_d(b_1) = 0$ and $I_d(b_2) = I_d(1/2)$, by Rolle's theorem, there exists $c_1 \in (0, b_1)$ and $c_2 \in (b_2, 1/2)$ such that $I'_d(c_1) = 0$ and $I'_d(c_2) = 0$.
- Since additionally $I'_d(1/2) = 0$, by Rolle's Theorem on $I'_d(x)$, there exists $d_1 \in (c_1, c_2)$ and $d_2 \in (c_2, 1/2)$ such that $I''_d(d_1) = 0$ and $I''_d(d_2) = 0$.

This means that $I''_d(x)$ has two zeros in $(0, 1/2)$, which contradicts Lemma 1.

REFERENCES

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.
- [2] A. E. Gamal and Y.-H. Kim, *Network Information Theory*. New York, NY, USA: Cambridge University Press, 2012.
- [3] C. Nair and A. E. Gamal, "An outer bound to the capacity region of the broadcast channel," *IEEE Transactions on Information Theory*, vol. 53, no. 1, pp. 350–355, Jan 2007.
- [4] C. Nair, "Capacity regions of two new classes of two-receiver broadcast channels," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4207–4214, Sept 2010.
- [5] A. A. Gohari and V. Anantharam, "Evaluation of marton's inner bound for the general broadcast channel," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 608–619, Feb 2012.
- [6] Y. Geng, C. Nair, S. S. Shitz, and Z. V. Wang, "On broadcast channels with binary inputs and symmetric outputs," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 6980–6989, Nov 2013.
- [7] Y. Geng, V. Jog, C. Nair, and Z. V. Wang, "An information inequality and evaluation of marton's inner bound for binary input broadcast channels," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4095–4105, July 2013.
- [8] A. Gohari, A. E. Gamal, and V. Anantharam, "On marton's inner bound for the general broadcast channel," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3748–3762, July 2014.
- [9] Y. Geng, A. Gohari, C. Nair, and Y. Yu, "On marton's inner bound and its optimality for classes of product broadcast channels," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 22–41, Jan 2014.
- [10] H. Kim, B. Nachman, and A. E. Gamal, "Superposition coding is almost always optimal for the Poisson broadcast channel," *IEEE Transactions on Information Theory*, vol. 62, no. 4, pp. 1782–1794, April 2016.
- [11] J. Korner and K. Marton, "A source network problem involving the comparison of two channels ii," in *Trans. Colloquium Inform. Theory, Keszthely, Hungary*, 1975.
- [12] A. E. Gamal, "The capacity of a class of broadcast channels," *IEEE Transactions on Information Theory*, vol. 25, no. 2, pp. 166–169, March 1979.
- [13] C. Nair, H. Kim, and A. E. Gamal, "On the optimality of randomized time division and superposition coding for the broadcast channel," in *2016 IEEE Information Theory Workshop (ITW)*, Sept 2016, pp. 131–135.
- [14] M. V. Dijk, "On a special class of broadcast channels with confidential messages," *IEEE Transactions on Information Theory*, vol. 43, no. 2, pp. 712–714, Mar 1997.

Maximally Recoverable Codes with Hierarchical Locality

Aaditya M. Nair, V. Lalitha

Abstract—Maximally recoverable codes are a class of codes which recover from all potentially recoverable erasure patterns given the locality constraints of the code. In earlier works, these codes have been studied in the context of codes with locality. The notion of locality has been extended to hierarchical locality, which allows for locality to gradually increase in levels with the increase in the number of erasures. We consider the locality constraints imposed by codes with two-level hierarchical locality and define maximally recoverable codes with data-local and local hierarchical locality. We derive certain properties related to their punctured codes and minimum distance. We give a procedure to construct hierarchical data-local MRCs from hierarchical local MRCs. We provide a construction of hierarchical local MRCs for all parameters. For the case of one global parity, we provide a different construction of hierarchical local MRC over a lower field size.

I. INTRODUCTION

With application to distributed storage systems, the notion of locality of a code was introduced in [1], which enables efficient node repair in case of single node failures (node failures modelled as erasures) by contacting fewer nodes than the conventional erasure codes based on maximum distance separable (MDS) codes. An extension to handle multiple erasures has been studied in [2]. A code symbol is said to have (r, δ) locality if there exists a punctured code \mathcal{C}_i such that $c_i \in \text{Supp}(\mathcal{C}_i)$ and the following conditions hold, 1) $|\text{Supp}(\mathcal{C}_i)| \leq r + \delta - 1$ and, 2) $d_{\min}(\mathcal{C}_i) \geq \delta$

An $[n, k, d_{\min}]$ code is said to have (r, δ) information locality, if k data symbols have (r, δ) locality and it is said to have all-symbol locality if all the n code symbols have (r, δ) locality. An upper bound on the minimum distance of a code with (r, δ) information locality is given by

$$d_{\min} \leq n - k + 1 - \left(\left\lceil \frac{k}{r} \right\rceil - 1 \right) (\delta - 1). \quad (1)$$

A. Maximally Recoverable Codes with Locality

Maximally recoverable codes (MRC) are a class of codes which recover from all information theoretically recoverable erasure patterns given the locality constraints of the code. Maximally recoverable codes with locality have been defined for the case of $\delta = 2$ in [3]. We extend the definitions here for the general δ .

Definition 1 (Data Local Maximally Recoverable Code). *Let C be a systematic $[n, k, d_{\min}]$ code. We say that C is*

Aaditya and Dr. Lalitha are with the Signal Processing & Communications Research Center, International Institute of Information Technology Hyderabad, India, email:aaditya.mnair@research.iiit.ac.in, lalitha.v@iiit.ac.in.

an $[k, r, h, \delta]$ data-local maximally recoverable code if the following conditions are satisfied

- $r|k$ and $n = k + \frac{k}{r}\delta + h$.
- Data symbols are partitioned into $\frac{k}{r}$ groups of size r . For each such group, there are δ local parity symbols.
- The remaining h global parity symbols may depend on all k symbols.
- For any set $E \subseteq [n]$ where E is obtained by picking δ coordinates from each $\frac{k}{r}$ local groups, restricting C to coordinates in $[n] - E$ yields a $[k + h, k]$ MDS code.

$[k, r, h, \delta]$ data-local MRC is optimum with respect to minimum distance bound in (1). The minimum distance of a $[k, r, h, \delta]$ data-local MRC is given by $d_{\min} = h + \delta + 1$.

Definition 2 (Local Maximally Recoverable Code). *Let C be a systematic $[n, k, d_{\min}]$ code. We say that C is an $[k, r, h, \delta]$ local maximally recoverable code if the following conditions are satisfied*

- $r|(k + h)$ and $n = k + \frac{k+h}{r}\delta + h$.
- There are k data symbols and h global parity symbols where each global parity may depend on all data symbols.
- These $k + h$ symbols are partitioned into $\frac{k+h}{r}$ groups of size r . For each group there are δ local parity symbols.
- For any set $E \subseteq [n]$ where E is obtained by picking δ coordinates from each $\frac{k+h}{r}$ local groups, restricting C to coordinates in $[n] - E$ yields a $[k + h, k]$ MDS code.

$[k, r, h, \delta]$ local MRC is optimum with respect to minimum distance bound in (1). The minimum distance of a $[k, r, h, \delta]$ local MRC is given by

$$d_{\min} = h + \delta + 1 + \left\lfloor \frac{h}{r} \right\rfloor \delta. \quad (2)$$

Maximally recoverable codes with locality for the case of general δ are known in literature as Partial-MDS codes (PMDS) codes. MRCs have been studied in the context of distributed storage systems and PMDS codes in the context of solid state drives (SSD) [4]. Constructions of PMDS codes with two and three global parities have been discussed in [5], [6]. A general construction of PMDS codes based on linearized polynomials has been provided in [7]. An improved construction of PMDS codes for all parameters over small field sizes ($\mathcal{O}(\max\{\frac{k+h}{r}, (r+\delta)^{\delta+h}\}^h)$) has been presented in [8]. Constructions of MRCs with field size $\mathcal{O}((\frac{k+h}{r})^r)$ have been presented in [12]. Construction of MRCs ($\delta = 2$) over small field sizes have been investigated in [9], [10].

B. Codes with Hierarchical Locality

The concept of *locality* has been extended to hierarchical locality in [11]. In the case of (r, δ) locality, if there are more than δ erasures, then the code offers no locality. In the case of codes with hierarchical locality, the locality constraints are such that with the increase in the number of erasures, the locality increases in steps. The following is the definition of code with two-level hierarchical locality.

Definition 3. An $[n, k, d_{min}]$ linear code \mathcal{C} is a code with hierarchical locality having parameters $[(r_1, \delta_1), (r_2, \delta_2)]$ if for every symbol c_i , $1 \leq i \leq n$, there exists a punctured code \mathcal{C}_i such that $c_i \in \text{Supp}(\mathcal{C}_i)$ and the following conditions hold, 1) $|\text{Supp}(\mathcal{C}_i)| \leq r_1 + \delta - 1$ 2) $d_{min}(\mathcal{C}_i) \geq \delta_1$ and 3) \mathcal{C}_i is a code with (r_2, δ_2) locality.

An upper bound on the minimum distance of a code with two-level hierarchical locality is given by

$$d \leq n - k + 1 - (\left\lceil \frac{k}{r_2} \right\rceil - 1)(\delta_2 - 1) - (\left\lceil \frac{k}{r_1} \right\rceil - 1)(\delta_1 - \delta_2). \quad (3)$$

C. Our Contributions

In this work, we consider the locality constraints imposed by codes with two-level hierarchical locality and define maximally recoverable codes with data-local and local hierarchical locality. We prove that certain punctured codes of these codes are data-local/local MRCs. We derive the minimum distance of hierarchical data-local MRCs. We give a procedure to construct hierarchical data-local MRCs from hierarchical local MRCs. We provide a construction of hierarchical local MRCs for all parameters. For the case of one global parity, we provide a different construction of hierarchical local MRC over a lower field size.

D. Notation

For any integer n , $[n] = \{1, 2, 3, \dots, n\}$. For any $E \subseteq [n]$, $\bar{E} = [n] - E$. For any $[n, k]$ code, and any $E \subseteq [n]$, $\mathcal{C}|_E$ refers to the punctured code obtained by restricting \mathcal{C} to the coordinates in E . This results in an $[n - |E|, k']$ code where $k' \leq k$. For any $m \times n$ matrix H and $E \subseteq [n]$, $H|_E$ is the $m \times |E|$ matrix formed by restricting H to columns indexed by E . In several definitions to follow, we implicitly assume certain divisibility conditions which will be clear from the context.

II. MAXIMALLY RECOVERABLE CODES WITH HIERARCHICAL LOCALITY

In this section, we define hierarchical data-local and local MRCs and illustrate the definitions through an example. We describe these codes via their parity check matrices instead of generator matrices (data local and local MRCs were defined by their generator matrices).

Definition 4 (Hierarchical Data Local Code). We define a $[k, r_1, r_2, h_1, h_2, \delta]$ hierarchical data local (HDL) code of length $n = k + h_1 + \frac{k}{r_1}(h_2 + \frac{r_1}{r_2}\delta)$ as follows:

- The code symbols c_1, \dots, c_n satisfy h_1 global parities given by $\sum_{j=1}^n u_j^{(\ell)} c_j = 0$, $1 \leq \ell \leq h_1$.

- The first $n - h_1$ code symbols are partitioned into $t_1 = \frac{k}{r_1}$ groups $A_i, 1 \leq i \leq t_1$ such that $|A_i| = r_1 + h_2 + \frac{r_1}{r_2}\delta = n_1$. The code symbols in the i^{th} group, $1 \leq i \leq t_1$ satisfy the following h_2 mid-level parities $\sum_{j=1}^{n_1} v_{i,j}^{(\ell)} c_{(i-1)n_1+j} = 0$, $1 \leq \ell \leq h_2$.
- The first $n_1 - h_2$ code symbols of the i^{th} group, $1 \leq i \leq t_1$ are partitioned into $t_2 = \frac{r_1}{r_2}$ groups $B_{i,s}, 1 \leq i \leq t_1, 1 \leq s \leq t_2$ such that $|B_{i,s}| = r_2 + \delta = n_2$. The code symbols in the $(i, s)^{\text{th}}$ group, $1 \leq i \leq t_1, 1 \leq s \leq t_2$ satisfy the following δ local parities $\sum_{j=1}^{n_2} w_{i,s,j}^{(\ell)} c_{(i-1)n_1+(s-1)n_2+j} = 0$, $1 \leq \ell \leq \delta$.

Definition 5 (Hierarchical Data Local MRC). Let \mathcal{C} be a $[k, r_1, r_2, h_1, h_2, \delta]$ HDL code. Then \mathcal{C} is maximally recoverable if for any set $E \subseteq [n]$ such that $|E| = k + h_1$, $|E \cap B_{i,s}| \geq r_2 \forall i, s$ and $|E \cap A_i| = r_1 \forall i$, the punctured code $\mathcal{C}|_E$ is a $[k + h_1, k, h_1 + 1]$ MDS code.

Definition 6 (Hierarchical Local Code). We define a $[k, r_1, r_2, h_1, h_2, \delta]$ hierarchical local (HL) code of length $n = k + h_1 + \frac{k+h_1}{r_1}(h_2 + \frac{r_1+h_2}{r_2}\delta)$ as follows:

- The code symbols c_1, \dots, c_n satisfy h_1 global parities given by $\sum_{j=1}^n u_j^{(\ell)} c_j = 0$, $1 \leq \ell \leq h_1$.
- The n code symbols are partitioned into $t_1 = \frac{k+h_1}{r_1}$ groups $A_i, 1 \leq i \leq t_1$ such that $|A_i| = r_1 + h_2 + \frac{r_1+h_2}{r_2}\delta = n_1$. The code symbols in the i^{th} group, $1 \leq i \leq t_1$ satisfy the following h_2 mid-level parities $\sum_{j=1}^{n_1} v_{i,j}^{(\ell)} c_{(i-1)n_1+j} = 0$, $1 \leq \ell \leq h_2$.
- The n_1 code symbols of the i^{th} group, $1 \leq i \leq t_1$ are partitioned into $t_2 = \frac{r_1+h_2}{r_2}$ groups $B_{i,s}, 1 \leq i \leq t_1, 1 \leq s \leq t_2$ such that $|B_{i,s}| = r_2 + \delta = n_2$. The code symbols in the $(i, s)^{\text{th}}$ group, $1 \leq i \leq t_1, 1 \leq s \leq t_2$ satisfy the following δ local parities $\sum_{j=1}^{n_2} w_{i,s,j}^{(\ell)} c_{(i-1)n_1+(s-1)n_2+j} = 0$, $1 \leq \ell \leq \delta$.

Definition 7 (Hierarchical Local MRC). Same as Definition 5.

In an independent parallel work [12], a class of MRCs known as multi-layer MRCs have been introduced. We would like to note that hierarchical local MRCs (given in Definition 7) form a subclass of these multi-layer MRCs.

Example 1. We demonstrate the structure of the parity check matrix for an $[k = 5, r_1 = 3, r_2 = 2, h_1 = 1, h_2 = 1, \delta = 2]$ HL code. The length of the code is $n = k + h_1 + \frac{k+h_1}{r_1}(h_2 + \frac{r_1+h_2}{r_2}\delta) = 16$. The parity check matrix of the code is given below:

$$H = \begin{bmatrix} M_{1,1} & & & \\ & M_{1,2} & & \\ \hline & N_1 & & \\ \hline & & M_{2,1} & \\ & & & M_{2,2} \\ \hline & & N_2 & \\ \hline & & & P \end{bmatrix}$$

$$M_{i,j} = \begin{bmatrix} w_{i,j,1}^{(1)} & w_{i,j,2}^{(1)} & w_{i,j,3}^{(1)} & w_{i,j,4}^{(1)} \\ w_{i,j,1}^{(2)} & w_{i,j,2}^{(2)} & w_{i,j,3}^{(2)} & w_{i,j,4}^{(2)} \end{bmatrix},$$

$$N_i = \begin{bmatrix} v_{i,1}^{(1)} & v_{i,2}^{(1)} & \dots & v_{i,8}^{(1)} \end{bmatrix} \text{ and } P = \begin{bmatrix} u_1^{(1)} & \dots & u_{16}^{(1)} \end{bmatrix}$$

III. PROPERTIES OF MRC WITH HIERARCHICAL LOCALITY

In this section, we will derive two properties of MRC with hierarchical locality. We will show that the middle codes of a HDL/HL-MRC have to be data-local and local MRC respectively. Also, we derive the minimum distance of HDL MRC.

Lemma III.1. Consider a $[k, r_1, r_2, h_1, h_2, \delta]$ HDL-MRC \mathcal{C} . Let $A_i, 1 \leq i \leq t_1$ be the supports of the middle codes as defined in Definition 4. Then, for each i , \mathcal{C}_{A_i} is a $[r_1, r_2, h_2, \delta]$ data-local MRC.

Proof. Suppose not. This means that for some i , the middle code \mathcal{C}_{A_i} is not a $[r_1, r_2, h_2, \delta]$ data-local MRC. By the definition of data-local MRC, we have that there exists a set $E_1 \subset A_i$ such that $|E_1| = r_1 + h_2$ and \mathcal{C}_{E_1} is not an $[r_1 + h_2, r_1, h_2 + 1]$ MDS code. This implies that there exists a subset $E' \subset E_1$ such that $|E'| = r_1$ and $\text{rank}(G|_{E'}) < r_1$. We can extend the set E' to obtain a set $E \subset [n]$, $|E| = k + h_1$ which satisfies the conditions in the definition of HDL-MRC. The resulting punctured code \mathcal{C}_E cannot be MDS since there exists an $r_1 < k$ sized subset of E such that $\text{rank}(G|_{E'}) < r_1$. \square

Lemma III.2. Consider a $[k, r_1, r_2, h_1, h_2, \delta]$ HL-MRC \mathcal{C} . Let $A_i, 1 \leq i \leq t_1$ be the supports of the middle codes as defined in Definition 6. Then, for each i , \mathcal{C}_{A_i} is a $[r_1, r_2, h_2, \delta]$ local MRC.

Proof. Proof is similar to the proof of Lemma III.1. \square

A. Minimum Distance of HDL-MRC

Lemma III.3. The minimum distance of a $[k, r_1, r_2, h_1, h_2, \delta]$ HDL-MRC is given by $d = h_1 + h_2 + \delta + 1$.

Proof. Based on the definition of HDL-MRC, it can be seen that the $[k, r_1, r_2, h_1, h_2, \delta]$ HDL-MRC is a code with hierarchical locality as per Definition 3 with k, r_1, r_2 being the same, $\delta_2 - 1 = \delta$, $\delta_1 = h_2 + \delta + 1$ and $n = k + h_1 + \frac{k}{r_1}(h_2 + \frac{r_1}{r_2}\delta)$. Substituting these parameters in the minimum distance bound in (3), we have that $d \leq h_1 + h_2 + \delta + 1$.

By Lemma III.1, we know that \mathcal{C}_{A_i} is a $[r_1, r_2, h_2, \delta]$ data-local MRC. The minimum distance of \mathcal{C}_{A_i} (from (2)) is $h_2 + \delta + 1$. Thus, the middle code itself can recover from any $h_2 + \delta$ erasures. The additional h_1 erasures can be shown to be extended to a set E (consisting of k additional non-erased symbols) which satisfies the conditions in Definition 5. Since, the punctured code $\mathcal{C}|_E$ is a $[k+h_1, k, h_1+1]$ MDS code, it can be used to recover the h_1 erasures. Hence, $[k, r_1, r_2, h_1, h_2, \delta]$ HDL-MRC can recover from any $h_1 + h_2 + \delta$ erasures. \square

B. Deriving HDL-MRC from HL-MRC

In this section, we give a method to derive any HDL-MRC from a HL-MRC. Assume an $[k, r_1, r_2, h_1, h_2, \delta]$ HL-MRC \mathcal{C} . Consider a particular set E of $k + h_1$ symbols satisfying the conditions given in Definition 7. We will refer to the elements

of set E as “primary symbols”. By the definition of HL-MRC, the code \mathcal{C} when punctured to E results in a $[k+h_1, k, h_1+1]$ MDS code. Hence, any k subset of E forms an information set. We will refer to the first k symbols of E as “data symbols” and the rest h_1 symbols as global parities. The symbols in $[n] \setminus E$ will be referred to as parity symbols (mid-level parities and local parities) and it can be observed that the parity symbols can be obtained as linear combinations of data symbols.

- If $r_1 \mid h_1$ and $r_2 \mid h_2$,
 - 1) For $A_i, \frac{k}{r_1} < i \leq \frac{k+h_1}{r_1}$, drop all the parity symbols, including h_2 mid-level parities per A_i as well as the δ local parities per $B_{i,s} \subset A_i$. As a result, we would be left with h_1 “primary symbols” in the local groups $A_i, \frac{k}{r_1} < i \leq \frac{k+h_1}{r_1}$. These form the global parities of the HDL-MRC. This step ensures that mid-level and local parities formed from global parities are dropped.
 - 2) For each $B_{i,s}, 1 \leq i \leq \frac{k}{r_1}, s > \frac{r_1}{r_2}$, drop the δ local parities. This step ensures that local parities formed from mid-level parities are dropped.

This results in an $[k, r_1, r_2, h_1, h_2, \delta]$ HDL-MRC.

- If $r_1 \nmid h_1$ and $r_2 \mid h_2$,
 - 1) From the groups $A_i, \lfloor \frac{k}{r_1} \rfloor + 1 < i \leq \frac{k+h_1}{r_1}$, drop all the parity symbols, including h_2 mid-level parities per A_i as well as the δ local parities per $B_{i,s} \subset A_i$.
 - 2) For each $B_{i,s}, 1 \leq i \leq \lfloor \frac{k}{r_1} \rfloor, s > \frac{r_1}{r_2}$, drop the δ local parities.
 - 3) Drop the $k - \lfloor \frac{k}{r_1} \rfloor r_1$ data symbols in $A_i, i = \lfloor \frac{k}{r_1} \rfloor + 1$ and recalculate all the parities (local, mid-level and global) by setting these data symbols as zero in the linear combinations.

This results in an $[\lfloor \frac{k}{r_1} \rfloor r_1, r_1, r_2, h_1, h_2, \delta]$ HDL-MRC.

For the case of $r_2 \nmid h_2$, HDL-MRC can be derived from HL-MRC using similar techniques as above. Hence, in the rest of the paper, we will discuss the construction of HL-MRC.

IV. GENERAL CONSTRUCTION OF HL-MRC

In this section, we will present a general construction of $[k, r_1, r_2, h_1, h_2, \delta]$ HL-MRC. First, we will provide the structure of the code and then derive necessary and sufficient conditions for the code to be HL-MRC. Finally, we will apply a known result of BCH codes to complete the construction.

Definition 8. A multiset $S \subseteq \mathbb{F}$ is k -wise independent over \mathbb{F} if for every set $T \subseteq S$ such that $|T| \leq k$, T is linearly independent over \mathbb{F} .

Lemma IV.1. Let \mathbb{F}_{q^t} be an extension of \mathbb{F}_q . Let a_1, a_2, \dots, a_n be elements of \mathbb{F}_{q^t} . The following matrix

$$\begin{bmatrix} a_1 & a_2 & a_3 & \dots & a_n \\ a_1^q & a_2^q & a_3^q & \dots & a_n^q \\ \vdots & \vdots & \vdots & \dots & \vdots \\ a_1^{q^{k-1}} & a_2^{q^{k-1}} & a_3^{q^{k-1}} & \dots & a_n^{q^{k-1}} \end{bmatrix}$$

is the generator matrix of a $[n, k]$ MDS code if and only if a_1, a_2, \dots, a_n are k -wise linearly independent over \mathbb{F}_q .

Proof. Directly follows from Lemma 3 in [8]. \square

Construction IV.2. The structure of the parity check matrix(H) of a $[k, r_1, r_2, h_1, h_2, \delta]$ HL-MRC is given by

$$H = \begin{bmatrix} H_0 & & & \\ & H_0 & & \\ & & \ddots & \\ & & & H_0 \\ H_1 & H_2 & \dots & H_{t_1} \end{bmatrix} H_0 = \begin{bmatrix} M_0 & & & \\ & M_0 & & \\ & & \ddots & \\ & & & M_0 \\ M_1 & M_2 & \dots & M_{t_2} \end{bmatrix}$$

Here, H_0 is an $(t_2\delta + h_2) \times n_1$ matrix and $H_i, 1 \leq i \leq t_1$ are an $h_1 \times n_1$ matrix. H_0 is then further subdivided into M_i . M_0 has the dimensions $\delta \times n_2$ and $M_i, 1 \leq i \leq t_2$ is an $h_2 \times n_2$ matrix.

Assume q to be a prime power such that $q \geq n$, $\mathbb{F}_{q^{m_1}}$ be an extension field of \mathbb{F}_q and \mathbb{F}_{q^m} is an extension field of $\mathbb{F}_{q^{m_1}}$, where $m_1 | m$.

In this case, the construction is given by the following.

$$M_0 = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 0 & \beta & \beta^2 & \dots & \beta^{n_2-1} \\ 0 & \beta^2 & \beta^4 & \dots & \beta^{2(n_2-1)} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & \beta^{\delta-1} & \beta^{2(\delta-1)} & \dots & \beta^{(\delta-1)(n_2-1)} \end{bmatrix},$$

where $\beta \in \mathbb{F}_q$ is a primitive element.

$$M_i = \begin{bmatrix} \alpha_{i,1} & \alpha_{i,2} & \dots & \alpha_{i,n_2} \\ \alpha_{i,1}^q & \alpha_{i,2}^q & \dots & \alpha_{i,n_2}^q \\ \vdots & \vdots & \dots & \vdots \\ \alpha_{i,1}^{q^{h_2-1}} & \alpha_{i,2}^{q^{h_2-1}} & \dots & \alpha_{i,n_2}^{q^{h_2-1}} \end{bmatrix},$$

where $i \in [t_2]$, $\alpha_{i,j} \in \mathbb{F}_{q^{m_1}}, 1 \leq i \leq t_2, 1 \leq j \leq n_2$.

$$H_i = [H_{i,1} \ H_{i,2} \ \dots \ H_{i,t_2}]$$

$$H_{i,s} = \begin{bmatrix} \lambda_{i,s,1} & \lambda_{i,s,2} & \dots & \lambda_{i,s,n_2} \\ \lambda_{i,s,1}^{q^{m_1}} & \lambda_{i,s,2}^{q^{m_1}} & \dots & \lambda_{i,s,n_2}^{q^{m_1}} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_{i,s,1}^{q^{m_1(h_1-1)}} & \lambda_{i,s,2}^{q^{m_1(h_1-1)}} & \dots & \lambda_{i,s,n_2}^{q^{m_1(h_1-1)}} \end{bmatrix},$$

where $i \in [t_1], s \in [t_2], \lambda_{i,s,j} \in \mathbb{F}_{q^m}, 1 \leq i \leq t_1, 1 \leq s \leq t_2, 1 \leq j \leq n_2$.

A (δ, h_2) erasure pattern is defined by the following two sets:

Δ is a three dimensional array of indices with the first dimension i indexing the middle code and hence $1 \leq i \leq t_1$, the second dimension s indexing the local code and hence $1 \leq s \leq t_2$. The third dimension j varies from 1 to δ and used to index the δ coordinates which are erased in the $(i, s)^{\text{th}}$ group. Let $e \in [n]$ denote the actual index of the erased coordinate in the code and $e \in B_{i,s}$, then we set $\Delta_{i,s,j} = (e \bmod n_2) + 1$. $\Delta_{i,s}$ is used to denote the vector of δ coordinates which are erased in the $(i, s)^{\text{th}}$ group. $\bar{\Delta}_{i,s}$ is used to denote the complement of $\Delta_{i,s}$ in the set $[n_2]$.

Γ is a two dimensional array of indices with the first dimension i indexing the middle code and hence $1 \leq i \leq t_1$. The second dimension j varies from 1 to h_2 and used to index the additional h_2 coordinates which are erased in the i^{th} group. Let $e \in [n]$ denote the actual index of the erased coordinate in the code and $e \in A_i$, then we set $\Gamma_{i,j} = (e \bmod n_1) + 1$. Γ_i is used to denote the vector of h_2 coordinates which are erased in the i^{th} group. $\bar{\Gamma}_i$ is used to denote the complement of Γ_i in the set $[n_1] \setminus (\cup_{s=1}^{t_2} \Delta_{i,s})$.

We define some matrices and sets based on the parameters of the construction, which will be useful in proving the subsequent necessary and sufficient condition for the construction to be HL-MRC. Here, $\alpha_{s,\Delta_{i,s}}$ denotes the set $\{\alpha_{s,j} \mid j \in \Delta_{i,s}\}$.

$$\begin{aligned} L_{i,s} &= (M_0|_{\Delta_{i,s}})^{-1} M_0|_{\bar{\Delta}_{i,s}} \\ \Psi_i &= \{\alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s}, 1 \leq s \leq t_2\} \\ &= \{\Psi_{i,\Gamma_i}, \Psi_{i,\bar{\Gamma}_i}\} \\ &= \{\psi_{i,1}, \dots, \psi_{i,h_2}, \psi_{i,h_2+1}, \dots, \psi_{i,r_1+h_2}\} \end{aligned}$$

The above equalities follow by noting that the $\cup_{s=1}^{t_2} \bar{\Delta}_{i,s} = \Gamma_i \cup \bar{\Gamma}_i$. We will refer to the elements in Ψ_{i,Γ_i} by $\{\psi_{i,1}, \dots, \psi_{i,h_2}\}$ and those in $\Psi_{i,\bar{\Gamma}_i}$ by $\{\psi_{i,h_2+1}, \dots, \psi_{i,r_1+h_2}\}$. Consider the following matrix based on the elements of Ψ_i ,

$$F_i = [F_i|_{\Gamma_i} \ F_i|_{\bar{\Gamma}_i}] = \begin{bmatrix} \psi_{i,1} & \psi_{i,2} & \dots & \psi_{i,r_1+h_2} \\ \psi_{i,1}^q & \psi_{i,2}^q & \dots & \psi_{i,r_1+h_2}^q \\ \vdots & \vdots & \dots & \vdots \\ \psi_{i,1}^{q^{h_2-1}} & \psi_{i,2}^{q^{h_2-1}} & \dots & \psi_{i,r_1+h_2}^{q^{h_2-1}} \end{bmatrix}, \quad (4)$$

And

$$\begin{aligned} \Phi_i &= \{\lambda_{i,s,\bar{\Delta}_{i,s}} + \lambda_{i,s,\Delta_{i,s}} L_{i,s}, 1 \leq s \leq t_2\} \\ &= \{\Phi_{i,\Gamma_i}, \Phi_{i,\bar{\Gamma}_i}\} \\ &= \{\phi_{i,1}, \dots, \phi_{i,h_2}, \phi_{i,h_2+1}, \dots, \phi_{i,r_1+h_2}\} \end{aligned}$$

Let $Z_i = (F_i|_{\Gamma_i})^{-1} F_i|_{\bar{\Gamma}_i}$. Finally, the set $\Theta = \{\Phi_{i,\bar{\Gamma}_i} + \Phi_{i,\Gamma_i} Z_i, 1 \leq i \leq t_1\}$.

Theorem IV.3. The code described in Construction IV.2 is a $[k, r_1, r_2, h_1, h_2, \delta]$ HL-MRC only if, for any (δ, h_2) erasure pattern, each $\Psi_i, 1 \leq i \leq t_1$ is h_2 -wise independent over \mathbb{F}_q and Θ is h_1 -wise independent over $\mathbb{F}_{q^{m_1}}$.

Proof. By Lemma III.2, we have that \mathcal{C} is a HL-MRC only if the $\mathcal{C}|_{A_i}$ is a $[r_1, r_2, h_2, \delta]$ local MRC. By the definition of local MRC, a code is a $[r_1, r_2, h_2, \delta]$ local MRC, if after puncturing δ coordinates in each of the $\frac{r_1+h_2}{r_2}$ local groups, the resultant code is $[r_1 + h_2, r_1, h_2 + 1]$ MDS code.

The puncturing on a set of coordinates in the code is equivalent to shortening on the same set of coordinates in the dual code. Shortening on a set of coordinates in the dual code can be performed by zeroing the corresponding coordinates in the parity check matrix by row reduction. To prove that $\mathcal{C}|_{A_i}$ is a $[r_1, r_2, h_2, \delta]$ local MRC, we need to show that certain punctured codes are MDS (Definition 2). We will equivalently that the shortened codes of the dual code are MDS.

Consider the coordinates corresponding to $(i, s)^{\text{th}}$ group in the parity check matrix. The sub-matrix of interest in this case is the following:

$$\left[\begin{array}{c|c} M_0|_{\Delta_{i,s}} & M_0|_{\bar{\Delta}_{i,s}} \\ \hline \alpha_{s,\Delta_{i,s}} & \alpha_{s,\bar{\Delta}_{i,s}} \\ \alpha_{s,\Delta_{i,s}}^q & \alpha_{s,\bar{\Delta}_{i,s}}^q \\ \vdots & \vdots \\ \alpha_{s,\Delta_{i,s}}^{q^{h_2-1}} & \alpha_{s,\bar{\Delta}_{i,s}}^{q^{h_2-1}} \end{array} \right],$$

Where $\alpha_{s,\Delta_{i,s}}^q$ is the vector obtained by taking q^{th} power of each element in the vector. Applying row reduction to the above matrix, we have

$$\left[\begin{array}{c|c} M_0|_{\Delta_{i,s}} & M_0|_{\bar{\Delta}_{i,s}} \\ \hline \mathbf{0} & \alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s} \\ \mathbf{0} & (\alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s})^q \\ \vdots & \vdots \\ \mathbf{0} & (\alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s})^{q^{h_2-1}} \end{array} \right].$$

Note that $L_{i,s}$ can be pushed into the power of q since the elements of $L_{i,s}$ are in \mathbb{F}_q . After row reducing δ coordinates from each of the $\frac{r_1+h_2}{r_2}$ local groups in A_i , the resultant parity check matrix is F_i . Applying Lemma IV.1, F_i forms the generator matrix of an MDS code if and only if the set Ψ_i is h_2 -wise independent over \mathbb{F}_q . The shortening of the code above is applicable to mid-level parities. Now, we will apply similar shortening in two steps to global parities. The sub-matrix of interest in this case is the following:

$$\left[\begin{array}{c|c} M_0|_{\Delta_{i,s}} & M_0|_{\bar{\Delta}_{i,s}} \\ \hline \alpha_{s,\Delta_{i,s}} & \alpha_{s,\bar{\Delta}_{i,s}} \\ \alpha_{s,\Delta_{i,s}}^q & \alpha_{s,\bar{\Delta}_{i,s}}^q \\ \vdots & \vdots \\ \alpha_{s,\Delta_{i,s}}^{q^{h_2-1}} & \alpha_{s,\bar{\Delta}_{i,s}}^{q^{h_2-1}} \\ \hline \lambda_{i,s,\Delta_{i,s}} & \lambda_{i,s,\bar{\Delta}_{i,s}} \\ \lambda_{i,s,\Delta_{i,s}}^{q^{m_1}} & \lambda_{i,s,\bar{\Delta}_{i,s}}^{q^{m_1}} \\ \vdots & \vdots \\ \lambda_{i,s,\Delta_{i,s}}^{q^{m_1(h_1-1)}} & \lambda_{i,s,\bar{\Delta}_{i,s}}^{q^{m_1(h_1-1)}} \end{array} \right]$$

Applying row reduction to the above matrix, we have

$$\left[\begin{array}{c|c} M_0|_{\Delta_{i,s}} & M_0|_{\bar{\Delta}_{i,s}} \\ \hline \mathbf{0} & \alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s} \\ \mathbf{0} & (\alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s})^q \\ \vdots & \vdots \\ \mathbf{0} & (\alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s})^{q^{h_2-1}} \\ \hline \mathbf{0} & \lambda_{i,s,\bar{\Delta}_{i,s}} + \lambda_{i,s,\Delta_{i,s}} L_{i,s} \\ \mathbf{0} & (\lambda_{i,s,\bar{\Delta}_{i,s}} + \lambda_{i,s,\Delta_{i,s}} L_{i,s})^{q^{m_1}} \\ \vdots & \vdots \\ \mathbf{0} & (\lambda_{i,s,\bar{\Delta}_{i,s}} + \lambda_{i,s,\Delta_{i,s}} L_{i,s})^{q^{m_1(h_1-1)}} \end{array} \right].$$

To apply row reduction again, we consider the following sub-

matrix obtained by deleting the zero columns and aggregating the non-zero columns from the $\frac{r_1+h_2}{r_2}$ groups,

$$\left[\begin{array}{c|c} F_i|_{\Gamma_i} & F_i|_{\bar{\Gamma}_i} \\ \hline \Phi_{i,\Gamma_i} & \Phi_{i,\bar{\Gamma}_i} \\ \Phi_{i,\Gamma_i}^{q^{m_1}} & \Phi_{i,\bar{\Gamma}_i}^{q^{m_1}} \\ \vdots & \vdots \\ \Phi_{i,\Gamma_i}^{q^{m_1(h_1-1)}} & \Phi_{i,\bar{\Gamma}_i}^{q^{m_1(h_1-1)}} \end{array} \right].$$

Applying row reduction to the above matrix, we have

$$\left[\begin{array}{c|c} F_i|_{\Gamma_i} & F_i|_{\bar{\Gamma}_i} \\ \hline \mathbf{0} & \Phi_{i,\bar{\Gamma}_i} + \Phi_{i,\Gamma_i} Z_i \\ \mathbf{0} & (\Phi_{i,\bar{\Gamma}_i} + \Phi_{i,\Gamma_i} Z_i)^{q^{m_1}} \\ \vdots & \vdots \\ \mathbf{0} & (\Phi_{i,\bar{\Gamma}_i} + \Phi_{i,\Gamma_i} Z_i)^{q^{m_1(h_1-1)}} \end{array} \right].$$

Note that Z_i can be pushed into the power of q^{m_1} since the elements of Z_i are in $\mathbb{F}_{q^{m_1}}$. Applying Lemma IV.1, the row reduced matrix above forms the generator matrix of an MDS code if and only if the set Θ is h_1 -wise independent over $\mathbb{F}_{q^{m_1}}$. \square

Lemma IV.4. *For any (δ, h_2) erasure pattern,*

- For each i , $\Psi_i = \{\alpha_{s,\bar{\Delta}_{i,s}} + \alpha_{s,\Delta_{i,s}} L_{i,s}, 1 \leq s \leq t_2\}$ is h_2 -wise independent over \mathbb{F}_q if the set $\{\alpha_{s,j}, 1 \leq s \leq t_2, 1 \leq j \leq n_2\}$ is $(\delta+1)h_2$ -wise independent over \mathbb{F}_q .
- $\Theta = \{\Phi_{i,\bar{\Gamma}_i} + \Phi_{i,\Gamma_i} Z_i, 1 \leq i \leq t_1\}$ is h_1 -wise independent over $\mathbb{F}_{q^{m_1}}$ if the set $\{\lambda_{i,s,j}, 1 \leq i \leq t_1, 1 \leq s \leq t_2, 1 \leq j \leq n_2\}$ is $(\delta+1)(h_2+1)h_1$ -wise independent over $\mathbb{F}_{q^{m_1}}$.

Proof. Since the size of matrix $L_{i,s}$ is $\delta \times (n_2 - \delta)$, each element of Ψ_i can be a \mathbb{F}_q -linear combination of atmost $\delta + 1$ different $\alpha_{s,j}$. Consider \mathbb{F}_q -linear combination of h_2 elements in Ψ_i . The linear combination will have at most $(\delta+1)h_2$ different $\alpha_{s,j}$. Thus, if the set $\{\alpha_{s,j}\}$ is $(\delta+1)h_2$ -wise independent over \mathbb{F}_q , then Ψ_i is h_2 -wise independent over \mathbb{F}_q . To prove the second part, we note that each element of Φ_i is a linear combination of at most $\delta + 1$ different $\lambda_{i,s,j}$. Since the size of the matrix Z_i is $h_2 \times (n_1 - h_2)$, each element of Θ can be a $\mathbb{F}_{q^{m_1}}$ -linear combination of atmost $(\delta+1)(h_2+1)$ different $\lambda_{i,s,j}$. Consider $\mathbb{F}_{q^{m_1}}$ -linear combination of h_1 elements in Θ . The linear combination will have at most $(\delta+1)(h_2+1)h_1$ different $\lambda_{i,s,j}$. Thus, if the set $\{\lambda_{i,s,j}\}$ is $(\delta+1)(h_2+1)h_1$ -wise independent over $\mathbb{F}_{q^{m_1}}$, then Θ is h_1 -wise independent over $\mathbb{F}_{q^{m_1}}$. \square

We will design the $\{\alpha_{s,j}\}$ and $\{\lambda_{i,s,j}\}$ based on the Lemma IV.4 so that the field size is minimum possible. We will pick these based on the following two properties:

- **Property 1:** The columns of parity check matrix of an $[n, k, d]$ linear code over \mathbb{F}_q can be interpreted as n elements over $\mathbb{F}_{q^{n-k}}$ which are $(d-1)$ -wise linear independent over \mathbb{F}_q .

- **Property 2:** There exists $[n = q^t - 1, k, d]$ BCH codes over \mathbb{F}_q [13], where the parameters are related as

$$n - k = 1 + \left\lceil \frac{q-1}{q}(d-2) \right\rceil \lceil \log_2(n) \rceil.$$

Theorem IV.5. *The code in Construction IV.2 is a $[k, r_1, r_2, h_1, h_2, \delta]$ HL-MRC if the parameters are picked as follows:*

- 1) q is the smallest prime power greater than n_2 .
- 2) m_1 is chosen based on the following relation: $m_1 = 1 + \left\lceil \frac{q-1}{q}((\delta+1)h_2 - 1) \right\rceil \lceil \log_q(n_2 t_2) \rceil$.
- 3) $n_2 t_2$ elements $\{\alpha_{s,j}\}$ over $\mathbb{F}_{q^{m_1}}$ are set to be the columns of parity check matrix of the BCH code over \mathbb{F}_q with parameters $[n = q^{\lceil \log_q(n_2 t_2) \rceil} - 1, q^{\lceil \log_q(n_2 t_2) \rceil} - 1 - m_1, (\delta+1)h_2 + 1]$.
- 4) m is chosen to be the smallest integer dividing m_1 based on the following relation: $m \geq 1 + \left\lceil \frac{q^{m_1}-1}{q^{m_1}}((\delta+1)(h_2+1)h_1 - 1) \right\rceil \lceil \log_{q^{m_1}}(n) \rceil$.
- 5) n elements $\{\lambda_{i,s,j}\}$ over \mathbb{F}_{q^m} are set to be the columns of parity check matrix of the BCH code over $\mathbb{F}_{q^{m_1}}$ with parameters $[n = q^{m_1 \lceil \log_{q^{m_1}}(n) \rceil} - 1, q^{m_1 \lceil \log_{q^{m_1}}(n) \rceil} - 1 - m, (\delta+1)(h_2+1)h_1 + 1]$.

Proof. The proof follows from Lemma IV.4 and Properties 1 and 2. \square

V. HL-MRC CONSTRUCTION FOR $h_1 = 1$

In this section, we present a construction of HL-MRC for the case when $h_1 = 1$ over a field size lower than that provided by Construction IV.2.

Construction V.1. *The structure of the parity check matrix for the present construction is the same as that given in Construction IV.2. In addition, the matrices M_0 and M_i , $1 \leq i \leq t_2$ also remain the same. We modify the matrix H_i , $1 \leq i \leq t_1$ as follows:*

$$H_i = \begin{bmatrix} \alpha_{1,1}^{q^{h_2}} & \alpha_{1,2}^{q^{h_2}} & \dots & \alpha_{t_2,n_2}^{q^{h_2}} \end{bmatrix},$$

where $\{\alpha_{s,j} \in \mathbb{F}_{q^{m_1}}, 1 \leq s \leq t_2, 1 \leq j \leq n_2\}$ are chosen to be $(\delta+1)(h_2+1)$ -wise independent over \mathbb{F}_q based on Theorem IV.5.

Theorem V.2. *The code C given by Construction V.1 is a $[k, r_1, r_2, h_1 = 1, h_2, \delta]$ HL-MRC.*

Proof. We show that H can be used to correct all erasure patterns defined in Definition 7. From the definition the code should recover from δ erasures per $B_{i,s}$, h_2 additional erasures per A_i and 1 more erasure anywhere in the entire code.

Now, with $h_1 = 1$, the last erasure can be part of one group. Thus, effectively the code should recover from $h_2 + 1$ erasures per group. Suppose that the last erasure is in the i^{th} group. The

submatrix of interest for the $(i, s)^{\text{th}}$ local group is

$$\left[\begin{array}{c|c} M_0|_{\Delta_{i,s}} & M_0|_{\bar{\Delta}_{i,s}} \\ \hline \alpha_{s,\Delta_{i,s}} & \alpha_{s,\bar{\Delta}_{i,s}} \\ \alpha_{s,\Delta_{i,s}}^q & \alpha_{s,\bar{\Delta}_{i,s}}^q \\ \vdots & \vdots \\ \alpha_{s,\Delta_{i,s}}^{q^{h_2-1}} & \alpha_{s,\bar{\Delta}_{i,s}}^{q^{h_2-1}} \\ \hline \alpha_{s,\Delta_{i,s}}^{q^{h_2}} & \alpha_{s,\bar{\Delta}_{i,s}}^{q^{h_2}} \\ \hline \alpha_{s,\Delta_{i,s}} & \alpha_{s,\bar{\Delta}_{i,s}} \end{array} \right].$$

Following the proof of Theorem IV.3 and performing row reduction of δ coordinates, the resultant matrix is

$$\left[\begin{array}{cccc} \psi_{i,1} & \psi_{i,2} & \dots & \psi_{i,r_1+h_2} \\ \psi_{i,1}^q & \psi_{i,2}^q & \dots & \psi_{i,r_1+h_2}^q \\ \vdots & \vdots & \dots & \vdots \\ \psi_{i,1}^{q^{h_2-1}} & \psi_{i,2}^{q^{h_2-1}} & \dots & \psi_{i,r_1+h_2}^{q^{h_2-1}} \\ \hline \psi_{i,1}^{q^{h_2}} & \psi_{i,2}^{q^{h_2}} & \dots & \psi_{i,r_1+h_2}^{q^{h_2}} \\ \hline \psi_{i,1} & \psi_{i,2} & \dots & \psi_{i,r_1+h_2} \end{array} \right].$$

Now, by Lemma IV.1, it is the generator matrix of an MDS code if and only if Ψ_i is $(h_2 + 1)$ -wise independent over \mathbb{F}_q . \square

ACKNOWLEDGMENT

This work was supported partly by the Early Career Research Award (ECR/2016/000954) from Science and Engineering Research Board (SERB) to V. Lalitha.

REFERENCES

- [1] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, “On the locality of codeword symbols,” *IEEE Transactions on Information Theory*, vol. 58, no. 11, pp. 6925–6934, 2012.
- [2] G. M. Kamath, N. Prakash, V. Lalitha, and P. V. Kumar, “Codes with local regeneration and erasure correction,” *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4637–4660, 2014.
- [3] P. Gopalan, C. Huang, B. Jenkins, and S. Yekhanin, “Explicit maximally recoverable codes with locality,” *IEEE Trans. Information Theory*, vol. 60, no. 9, pp. 5245–5256, 2014.
- [4] M. Blaum, J. L. Hafner, and S. Hetzler, “Partial-mds codes and their application to raid type of architectures.” *IEEE Trans. Information Theory*, vol. 59, no. 7, pp. 4510–4519, 2013.
- [5] M. Blaum, J. S. Plank, M. Schwartz, and E. Yaakobi, “Construction of partial mds and sector-disk codes with two global parity symbols,” *IEEE Transactions on Information Theory*, vol. 62, no. 5, pp. 2673–2681, 2016.
- [6] J. Chen, K. W. Shum, Q. Yu, and C. W. Sung, “Sector-disk codes and partial mds codes with up to three global parities,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*, pp. 1876–1880, IEEE, 2015.
- [7] G. Calis and O. O. Koyleoglu, “A general construction for pmds codes,” *IEEE Communications Letters*, vol. 21, no. 3, pp. 452–455, 2017.
- [8] R. Gabrys, E. Yaakobi, M. Blaum, and P. H. Siegel, “Constructions of partial mds codes over small fields,” in *Information Theory (ISIT), 2017 IEEE International Symposium on*, pp. 1–5, IEEE, 2017.
- [9] G. Hu and S. Yekhanin, “New constructions of sd and mr codes over small finite fields,” *arXiv preprint arXiv:1605.02290*, 2016.
- [10] V. Guruswami, L. Jin, and C. Xing, “Constructions of maximally recoverable local reconstruction codes via function fields,” *arXiv preprint arXiv:1808.04539*, 2018.
- [11] B. Sasidharan, G. K. Agarwal, and P. V. Kumar, “Codes with hierarchical locality,” in *Information Theory (ISIT), 2015 IEEE International Symposium on*, pp. 1257–1261, IEEE, 2015.
- [12] U. Martínez-Peña and F. R. Kschischang, “Universal and dynamic locally repairable codes with maximal recoverability via sum-rank codes,” *arXiv preprint arXiv:1809.11158*, 2018.
- [13] R. Roth, *Introduction to coding theory*. Cambridge University Press, 2006.

Probability Mass Functions for which Sources have the Maximum Minimum Expected Length

Shivkumar K. Manickam

Dept. of Electrical Communication Engineering

Indian Institute of Science, Bangalore, India

Email: shivkumar@iisc.ac.in

Abstract—Let \mathcal{P}_n be the set of all probability mass functions (PMFs) (p_1, p_2, \dots, p_n) that satisfy $p_i > 0$ for $1 \leq i \leq n$. Define the minimum expected length function $\mathcal{L}_D : \mathcal{P}_n \rightarrow \mathbb{R}$ such that $\mathcal{L}_D(P)$ is the minimum expected length of a prefix code, formed out of an alphabet of size D , for the discrete memoryless source having P as its source distribution. It is well-known that the function \mathcal{L}_D attains its maximum value at the uniform distribution. Further, when n is of the form D^m , with m being a positive integer, PMFs other than the uniform distribution at which \mathcal{L}_D attains its maximum value are known. However, a complete characterization of all such PMFs at which the minimum expected length function attains its maximum value has not been done so far. This is done in this paper.

I. INTRODUCTION TO THE PROBLEM

One of the earliest problems considered in information theory is that of finding a prefix code with the minimum expected length for a given discrete memoryless source. This paper addresses a question related to this problem.

Let us begin by establishing the basic terminology and notation used in this paper. A set of finite length strings of letters coming from a given finite alphabet is said to be a *prefix code* if no string is a prefix of another. Let us use \mathcal{A} to denote the finite alphabet and D to denote its size. The members of a prefix code are called *codewords*.

Consider a discrete source with n symbols where the i th symbol occurs with a probability p_i ($\sum_i p_i = 1$). The collection of probabilities $P = (p_1, p_2, \dots, p_n)$ is a *probability mass function* (PMF). Let X be a prefix code assigned to this source (from now onwards, we will simply say X is a prefix code for the PMF P , leaving out any mention of the source) and let the i th source symbol be associated with a codeword of length l_i . The *expected length* of the prefix code X is $\sum_i p_i l_i$. A minimum expected length prefix code can be effectively obtained using the Huffman algorithm [1]. Henceforth, we will refer to a minimum expected length prefix code as an *optimal code*.

Let \mathcal{P}_n be the set of all PMFs (p_1, p_2, \dots, p_n) that satisfy $p_i > 0$ for $1 \leq i \leq n$. Define the function $\mathcal{L}_D : \mathcal{P}_n \rightarrow \mathbb{R}$ such that $\mathcal{L}_D(P)$ is the expected length of an optimal code for the PMF P . Let us call this function the *minimum expected length function*.

Now, for a PMF $P \in \mathcal{P}_n$, the only known general way to determine $\mathcal{L}_D(P)$ is by first determining an optimal code using the Huffman algorithm and then finding its expected length. There is neither any known analytical formula for $\mathcal{L}_D(P)$ in terms of the probabilities of P nor an alternate characterization of the function \mathcal{L}_D , from which its values can be readily evaluated. However, some properties of this function are known. Let us call a PMF at which \mathcal{L}_D attains its maximum value to be a *point of maximum*. A result of Hwang [2] shows that the minimum expected length function is Schur-concave, and so attains its maximum value at the uniform distribution: $U_n = (1/n, 1/n, \dots, 1/n)$. Further, when n is of the form D^m , a result of [3] gives other points of maximum: all $P \in \mathcal{P}_n$ in which the sum of the lowest D probabilities is greater than or equal to the highest probability (we will see more about this). However, to the author's knowledge, a complete characterization of all the points of maximum has not been done so far. This is carried out in this paper.

We will be making use of a characterization of Huffman trees given by Gallager [4]. This is presented in the next section.

II. HUFFMAN TREES

It is useful to visualize prefix codes in the form of trees [5, Chapter 5] (see Fig. 1). For our purpose, we will find it convenient to give directions to the edges of a tree. When we will refer to a directed graph as a tree, we will do so in the sense that the undirected graph obtained by replacing each directed edge by an undirected one is a tree. Consider an infinite rooted directed tree in which all the edges are directed away from the root. Let each node of the tree have D outgoing edges. Let us denote this tree as T_∞ . Throughout this paper, whenever we talk of a tree we will mean a subtree of T_∞ . In a tree, if there is an edge from a node v to node v_1 , then v is said to be the *parent* of v_1 and v_1 is said to be a *child* of v . A node d is said to be a *descendant* of a node v if there is a path from v to d . Nodes having the same parent are called *siblings*. A *sibling set* is the set of the children of an internal node of a tree. The *level* of a node v is the length of the path from the root to v .

For every node of T_∞ , label each of the outgoing edges with a letter from \mathcal{A} such that no two edges is associated

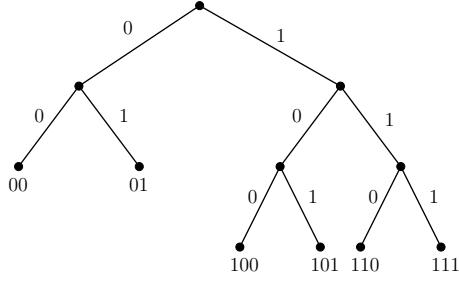


Fig. 1. Code tree for the prefix-free code $\{00, 01, 100, 101, 110, 111\}$.

with the same letter. Associate each node with the string formed by reading out the labels in the path from the root to that particular node. To get the tree representation of a prefix code, mark all the nodes of the labeled tree T_∞ that correspond to the codewords. Retain all the nodes in the paths from the root to the marked nodes (including the root and the marked nodes) and delete the remaining nodes. The resulting tree is called a *code tree*. Conversely, any tree T can be considered to represent a class of prefix codes. To get this class, consider all the possible ways the edges of the tree T can be labeled using the letters of \mathcal{A} (with no two outgoing edges from a node getting the same label). From each assignment of labels on the edges, we can get a prefix code by collecting all the strings along the leaves of T . Let us consider T to represent the class of all such prefix codes obtained from all the different assignments of labels.

Let T be a tree with n leaves and let $P \in \mathcal{P}_n$. If there is a 1-1 mapping between the probabilities of P and the leaves of T , we say that T is a tree associated with the PMF P . Using our relation between a tree and a class of prefix codes, we can see that a tree T associated with a PMF P defines a class of prefix codes for P . Observe that the expected length of all these prefix codes is the same, and we take this value as the *expected length* of the tree T . An *optimal tree* is that having the minimum expected length. The Huffman algorithm can be viewed as a one that constructs an optimal tree. It is briefly described below as we will refer to it to establish the Huffman tree characterization.

A. The Huffman Algorithm

Let $P = (p_1, p_2, \dots, p_n)$ be the given PMF for which we need an optimal code. Consider a forest F_1 containing n isolated vertices. View each of the vertices as a tree and associate the probability p_i with the i th vertex. Identify the integer m such that $2 \leq m \leq D$ and $D - 1$ divides $n - m$. Choose any m vertices v_1, v_2, \dots, v_m from F_1 having the lowest m probabilities. Add a new vertex r to F_1 and make it the parent of the vertices v_1, v_2, \dots, v_m . Associate with the vertex r the probability equaling the sum of the probabilities of its children. Finally, make the vertex r the root of the tree to which it belongs.

Now suppose that the forests F_1, F_2, \dots, F_i are defined, with F_i not being a tree. Define the forest F_{i+1} as follows: choose any D roots of the trees in F_i with lowest D probabilities. As in the previous case assign a parent to them, making the parent the root of the tree to which it belongs and assigning it the probability equaling the sum of the probabilities of its children. This is the forest F_{i+1} . From the way the first step of the algorithm was carried out, we will end with a forest that is in fact a tree associated with P ; let us call it a *Huffman tree* for the PMF P .

B. Huffman Tree Characterization

Let T be a tree for a PMF P . As done during the Huffman algorithm, let us associate each node of T with a probability in the following way: a node gets the probability equal to the sum of the probabilities of its children. This assignment will result in the root node getting the probability 1. It can be shown that the expected length of T is the sum of the probabilities of all its internal nodes (including the root node).

Gallager [4] has given a necessary and sufficient condition for a tree associated with a PMF to be a *Huffman tree*, i.e. a tree which can be generated by the application of Huffman algorithm on P . For our purpose, we will need a more descriptive version of this condition which is stated in the following theorem. A part of its proof follows the arguments presented in [4].

Theorem 1. *Let T be a tree associated with a PMF $P \in \mathcal{P}_n$. It is a Huffman tree for P iff the following conditions hold:*

- (P1) *The probability of a lower level node is greater than or equal to that of a higher level node.*
- (P2) *Let m be such that $D - 1$ divides $n - m$ with $2 \leq m \leq D$. The tree T contains a sibling set with m nodes at its maximum level which has the lowest m probabilities of P . All the other sibling sets of T have exactly D nodes.*
- (P3) *The nodes at any level of T can be listed in such a way that their probabilities are in a non-decreasing order, and the siblings come next to each other in the listing.*

Proof. Let us first take T to be a Huffman tree for P . Now suppose that the condition (P1) is not satisfied for two nodes v_1 and v_2 at levels l_1 and l_2 respectively, with $l_1 < l_2$. Let v_1 have the probability p_1 and v_2 have the probability p_2 . Let $ST(v_1)$ ($ST(v_2)$) be the subtree of T that has v_1 (v_2) as its root and contains all the descendants of v_1 (v_2). Let $l_{11}, l_{12}, \dots, l_{1i}$ be the levels of the leaves of $ST(v_1)$, with levels calculated from the node v_1 , and $p_{11}, p_{12}, \dots, p_{1i}$ be their respective probabilities (with $\sum_k p_{1k} = p_1$). Similarly, let $l_{21}, l_{22}, \dots, l_{2j}$ be the levels of the leaves of $ST(v_2)$, with levels calculated from the node v_2 , and $p_{21}, p_{22}, \dots, p_{2j}$ be their respective probabilities (with $\sum_k p_{2k} = p_2$). Since $p_2 > p_1$, v_2 cannot be a descendant of v_1 . Create a new tree associated with P as follows without changing the association between the

probabilities of P and the leaves of T : Let u_1 and u_2 be the parents of v_1 and v_2 respectively. Delete the edges between u_1 and v_1 , and u_2 and v_2 . Construct new edges from u_1 to v_2 , and u_2 to v_1 . In other words, we are interchanging the parents of v_1 and v_2 . It is clear that the resulting graph is a tree. Let us call this tree T' . Let $L(T)$ and $L(T')$ denote the expected lengths of T and T' respectively. Now, there exists a $\lambda \in \mathbb{R}$ such that

$$\begin{aligned} L(T) &= \lambda + \sum_{k=1}^i (l_1 + l_{1k})p_{1k} + \sum_{k=1}^j (l_2 + l_{2k})p_{2k} \\ &= \lambda + l_1 p_1 + l_2 p_2 + \sum_{k=1}^i l_{1k} p_{1k} + \sum_{k=1}^j l_{2k} p_{2k}, \end{aligned} \quad (1)$$

and

$$\begin{aligned} L(T') &= \lambda + \sum_{k=1}^i (l_1 + l_{2k})p_{2k} + \sum_{k=1}^j (l_2 + l_{1k})p_{1k} \\ &= \lambda + l_1 p_2 + l_2 p_1 + \sum_{k=1}^i l_{1k} p_{1k} + \sum_{k=1}^j l_{2k} p_{2k}. \end{aligned} \quad (2)$$

Notice that $L(T') < L(T)$ which contradicts the optimality of T . Thus the condition (P1) is true.

The condition (P2) follows from the way the Huffman algorithm is carried out and the condition (P1).

Let us now prove condition (P3). Let F_1, F_2, \dots, F_N be the forests obtained during the execution of the Huffman algorithm on PMF P that yields the tree T (see Section II-A). Let S_i ($1 \leq i \leq N$) be the chosen root nodes in F_i having the lowest D probabilities (lowest m probabilities when $i = 1$). It can be seen that the S_i 's are precisely the sibling sets in T . Further, each of the probabilities in S_i is less than or equal to each of that in S_{i+1} . Thus, it is possible to list all the nodes of T in such a way that their probabilities are in a non-decreasing order, and the siblings come next to each other in the listing. As a result, (P3) also holds.

Let us prove that these conditions are sufficient for T to be a Huffman tree for P . We will do so by showing that a tree isomorphic to T can be obtained by the application of the Huffman algorithm on P .

Let V and V_L denote the nodes and leaf nodes of T respectively. Let U be a set with $|U| = |V|$. Define a bijection $\psi : V \rightarrow U$. For any $V' \subseteq V$, let the ψ -image of V' be the set $\{\psi(v') \mid v' \in V'\}$. Construct a forest F exactly containing all the elements of the ψ -image of V_L as isolated vertices (with each of them viewed as a tree with that vertex itself serving as its root). For each node of V_L , assign its probability to its ψ -image in the forest. Let S be the sibling set in T having the m lowest probabilities of P , and let r be its parent. Derive a new forest F' from F by introducing $\psi(r)$ as a new vertex to F , and making it the parent of the vertices occurring in the ψ -image of S . Assign the probability of the node r to $\psi(r)$. Derive a new tree T' from T by deleting all the nodes of S from T . It

can be seen that the roots of the trees in F' are precisely the leaf nodes of T' .

If F' is not a tree, then perform the above mentioned step for T' and F' by taking S to be a sibling set in T' containing the lowest D probabilities of T' . Such a choice is possible as T' also satisfies conditions (P1) and (P3). Note that all the nodes of S are leaf nodes, for otherwise, we would have a node in T' with probability strictly less than that of a node in S . The execution of the step will leave us with a tree and a graph both of which are derived in two steps from T and F respectively. The graph derived from F will be a forest as the ψ -image of S are root nodes in F' . Keep repeating the above step and it can be seen that each execution of the step will yield a tree derived from T and a forest derived from F with the ψ -image of the leaves of the tree being exactly the root nodes of the trees of the forest. Continue doing so till the forest derived from F becomes a tree, say T^* . It can be seen that these steps constitute the Huffman algorithm and so T^* is a Huffman tree for P . By way of construction of T^* , it is clear that it is isomorphic to T under the bijection ψ . Thus, T is a Huffman tree for P . \square

Remark 1. Equations (1) and (2) can be used to show that if a tree T is optimal, then it should at least satisfy the condition (P1).

Let us call the conditions (P1)–(P3) as *Huffman tree properties*. We will be referring to each of them as follows: the condition (P1) will be called the *level property*, condition (P2) will be called the *maximum-level sibling property* and condition (P3) will be called the *sibling property*. These properties serve as a potent tool to approach questions related to Huffman trees.

III. POINTS OF MAXIMUM

Let the sequence of the codeword lengths of a prefix code X arranged in a non-decreasing order be called the *length sequence* of X . Let an *optimal length sequence* for a PMF P be the length sequence of an optimal code for P . Let us follow the convention of always writing out the probabilities of a PMF $P = (p_1, p_2, \dots, p_n)$ in a non-increasing order, i.e., $p_1 \geq p_2 \geq \dots \geq p_n$.

Let us now take up the following problem: what is an optimal length sequence for a point of maximum? The length sequence of the Huffman code for the uniform distribution — which is a point of maximum — is known (see, for e.g., [3]). To emphasize the point that problems related to Huffman trees can be effectively handled using the Huffman tree properties, we will now use these to determine the length sequence of the Huffman code for U_n .

Consider the tree having D^m leaves at level m , for some $m \in \mathbb{N}$. Let us denote it as $T_U(m)$. We have the following result which has appeared in [3].

Lemma 1. Let $n = D^m$, for some $m \in \mathbb{N}$. A PMF $P \in \mathcal{P}_n$ with $P = (p_1, p_2, \dots, p_n)$ has $T_U(m)$ as a Huffman tree iff

the sum of the lowest D probabilities of P is greater than or equal to its highest probability, i.e., iff

$$\sum_{i=n-D+1}^n p_i \geq p_1.$$

Proof. First suppose that $T_U(m)$ is a Huffman tree for P . The minimum probability of the nodes at level $m - 1$ is $\sum_{i=n-D+1}^n p_i$ (maximum-level sibling property (P2)). From the level property (P1), we have that $\sum_{i=n-D+1}^n p_i \geq p_1$.

Now suppose that $\sum_{i=n-D+1}^n p_i \geq p_1$. Assign the probabilities of P to the leaves of $T_U(m)$ in such a way that the probabilities of the leaves from left to right are in a non-decreasing order. We will now show $T_U(m)$ with this assignment of probabilities satisfies the Huffman tree properties. We will do it by induction on m . The statement is clearly true for $m = 1$. Let us assume that for some positive integer k , the statement is true for $m = k$. Let us now take $m = k + 1$. Let us denote the PMF formed out of the probabilities at level k of $T_U(k + 1)$ to be Q . Observe that the probabilities of Q are arranged in a non-decreasing order from left to right at the k th level. Let q_1, q_2, \dots, q_D be the lowest D probabilities of Q taken in a non-decreasing order. The highest probability of Q , say q_h , is given by $p_1 + p_2 + \dots + p_D$. Since $q_1 = \sum_{i=n-D+1}^n p_i$, we have the following inequalities:

$$q_D \geq q_{D-1} \geq \dots \geq q_1 \geq p_1 \geq p_2 \geq \dots \geq p_D.$$

Thus, we have that $\sum_{i=1}^D q_i \geq q_h$. By the induction hypothesis, we have that the tree $T_U(k)$, with its nodes retaining its probabilities as in $T_U(k+1)$, satisfies the Huffman tree properties. Now take a node at level k in the tree $T_U(k+1)$ with probability q_i and a node at level $k+1$ with probability p_j . We have that $q_i \geq q_1 \geq p_1 \geq p_j$. Thus, the level property (P1) is satisfied in $T_U(k+1)$. Finally, by the way the probabilities were assigned to the leaves of $T_U(k+1)$, the maximum-level sibling property (P2) and the sibling property (P3) are also satisfied at level $k+1$ of $T_U(k+1)$. Thus, from Theorem 1, the tree $T_U(k+1)$ is a Huffman tree for P . Hence the lemma is proved. \square

Remark 2. For the case $n = D^m$ ($m \in \mathbb{N}$), Lemma 1 describes points of maximum, other than the uniform distribution U_n , for the minimum expected length function \mathcal{L}_D . For a PMF $P \in \mathcal{P}_n$, satisfying the condition of Lemma 1, i.e., $\sum_{i=n-D+1}^n p_i \geq p_1$, we have from this lemma that $\mathcal{L}_D(P) = m$. Since U_n also satisfies this condition of the lemma, and since it is a point of maximum, we have that P is also a point of maximum. Thus, all the PMFs satisfying the condition of Lemma 1 are points of maximum.

Now, let T_U represent the tree $T_U(m)$ when $n = D^m$ (a power of D) and when $D^m < n < D^{m+1}$ (n is not a power of D) let T_U denote the tree (see Fig. 2) in which

- i) all the leaves are at levels m and $m + 1$,

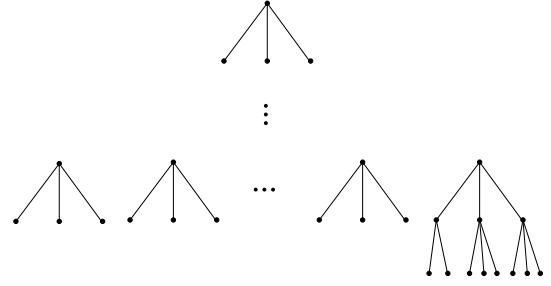


Fig. 2. An example for T_U when $D = 3$ and n is not a power of D .

- ii) at level m , each of the internal nodes is to the right of any of the leaf nodes and
 - iii) all the internal nodes, except possibly the leftmost internal node at level m , have D children each. The leftmost internal node at level m has at least 2 children.

It can be seen that these conditions uniquely define the tree T_U when n is not a power of D . When $D - 1$ divides $n - 1$, then all the internal nodes of T_U will have D children; and the number of internal nodes at level m is $(n - D^m)/(D - 1)$. When $D - 1$ doesn't divide $n - 1$, only one internal node will have m children where m is such that $2 \leq m \leq D$ and $D - 1$ divides $n - m$; and the number of internal nodes at level m is $\lceil (n - D^m)/(D - 1) \rceil$.

Now we have the following result.

Theorem 2. The tree T_U is a Huffman tree for the uniform distribution U_n .

Proof. Lemma 1 tells us that the theorem is true when n is a power of D . Let us take that n satisfies $D^m < n < D^{m+1}$. Consider T_U and assign the probabilities of U_n to its leaves. Consider the tree $T_U(m)$ obtained by removing all the leaves from T_U at level $m + 1$. Let the nodes of $T_U(m)$ retain their probabilities as they were in T_U . Now look at the probabilities of the nodes of $T_U(m)$ at level m . The lowest probability is at least $1/n$ and the highest probability is at most D/n . Thus, by Lemma 1 and Theorem 1, we have that $T_U(m)$ satisfies the Huffman tree properties. This, along with the way the tree T_U is defined shows us that T_U also satisfies the Huffman tree properties. Thus, T_U is a Huffman tree for U_n . \square

Thus, we have that when $n = D^m$, an optimal length sequence for U_n is (m, m, \dots, m) and when $D^m < n < D^{m+1}$ an optimal length sequence for U_n is $(m, m, \dots, m, m + 1, m + 1, \dots, m + 1)$ with $D^m - \lceil (n - D^m)/(D - 1) \rceil$ occurrences of m . Let us denote this optimal length sequence for U_n by L_U .

It turns out that Hwang's argument, as in [2], can now be used to determine an optimal length sequence for any point of maximum. Let $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ be two PMFs from \mathcal{P}_n . If the sum of the highest k probabilities of P is greater than or equal to the sum of the highest k probabilities of Q , i.e. if

$\sum_{i=1}^k p_i \geq \sum_{i=1}^k q_i$, for all k satisfying $1 \leq k \leq n$, then P is said to *majorize* Q and is denoted as $P \succ Q$. Note that this relation is a partial order. It is well-known that every $P \in \mathcal{P}_n$ majorizes the uniform distribution U_n .

Lemma 2. *If P is a point of maximum, then the following hold:*

- i) L_U is an optimal length sequence for P .
- ii) When n is not a power of D , the PMF P is the uniform distribution U_n , or, in other words, U_n is the unique point of maximum.

Proof. Let $P = (p_1, p_2, \dots, p_n)$ and let $\mathsf{L}_P = (l_1^{(p)}, l_2^{(p)}, \dots, l_n^{(p)})$ be an optimal length sequence for P . From our ordering convention, we have that $l_i^{(p)}$ is the length of the codeword associated with p_i . Let us also consider the optimal length sequence L_U for U_n and write it out as $(l_1^{(u)}, l_2^{(u)}, \dots, l_n^{(u)})$. Take $l_0^{(u)} = l_0^{(p)} = 0$. Making use of Hwang's technique, we get the following chain of inequalities:

$$\begin{aligned} \mathcal{L}_D(U_n) &= \sum_{i=1}^n l_i^{(u)} / n, \\ &= \sum_{i=1}^n (l_i^{(u)} - l_{i-1}^{(u)}) \sum_{j=i}^n \frac{1}{n}, \end{aligned} \quad (3)$$

$$\geq \sum_{i=1}^n (l_i^{(u)} - l_{i-1}^{(u)}) \sum_{j=i}^n p_j, \quad (4)$$

$$\begin{aligned} &= \sum_{i=1}^n l_i^{(u)} p_i, \\ &\geq \sum_{i=1}^n l_i^{(p)} p_i, \end{aligned} \quad (5)$$

$$= \mathcal{L}_D(P),$$

where (4) follows from the relation $P \succ U_n$, and (5) follows from the fact that L_P is an optimal length sequence for P . Since $\mathcal{L}_D(U_n) = \mathcal{L}_D(P)$, we have that the inequalities in (4) and in (5) are equalities. Let us see what they imply.

i) Since the inequality in (5) is actually an equality, we have that L_U is an optimal length sequence for P .

ii) Let n be such that $D^m < n < D^{m+1}$. Since the entries in the sequence L_U are either m and $m+1$, the expression in (3) boils down to $m \sum_{j=1}^n 1/n + \sum_{j=k+1}^n 1/n$, for k such that $l_k^{(u)} = m$ and $l_{k+1}^{(u)} = m+1$. Similarly, the expression in (4) is $m \sum_{j=1}^n p_j + \sum_{j=k+1}^n p_j$. Since the inequality in (4) is an equality, we have that

$$\frac{n-k}{n} = \sum_{j=k+1}^n p_j. \quad (6)$$

Let $I_1 = \{1, 2, \dots, k\}$ and $I_2 = \{k+1, k+2, \dots, n\}$. We have that p_{k+1} is greater than or equal to the average of $\{p_j\}_{j \in I_2}$, which equals $1/n$ (from 6). Thus, $p_{k+1} \geq 1/n$. From (6), we also have that

$$\sum_{j \in I_1} p_j = k/n. \quad (7)$$

For $j \in I_1$, since $p_j \geq p_{k+1} \geq 1/n$, (7) can occur iff

$$p_j = 1/n, \text{ for all } j \in I_1. \quad (8)$$

Now, if there exists a $j \in I_2$ such that $p_j < 1/n$, then (6) implies that $p_{k+1} > 1/n$. But this contradicts (8). Thus, P is the uniform distribution. Hence, when n is not a power of D , uniform distribution U_n is the unique point of maximum. \square

Lemma 2 tells that when n is not a power of D , the uniform distribution is the unique point of maximum. It also tells that when n is a power of D , the set of all the points of maximum is precisely the collection of all the PMFs that have L_U as their optimal length sequence. Is this set the same as the set of all PMFs whose Huffman code's length sequence is L_U ? The answer to this question is not clear at this stage because the Huffman algorithm does not generate all the optimal codes. The following result will come to our aid:

Lemma 3. *If L is an optimal length sequence for a $P \in \mathcal{P}_n$, then there exists a Huffman tree for P with L as its length sequence.*

Proof. Let T be an optimal tree for P with L as its length sequence. It should at least satisfy the level property (P1) (Remark 1).

Let l_{\max} be the maximum level of a node of T . Change, if necessary, the assignment of children at level l_{\max} to the internal nodes at level $l_{\max} - 1$ so that the resulting tree is still a subtree of T_∞ , but now the maximum-level sibling property (P2) and the sibling property (P3) are satisfied by the nodes at level l_{\max} . This will not create any new leaf nodes at level $l_{\max} - 1$ of the resulting tree, call it T' , for otherwise one of the leaves at level l_{\max} can be deleted and its probability can be assigned to one of the newly created leaves at level $l_{\max} - 1$. If we now throw away the other leaves, if present, without any assignment of probabilities, then the expected length of this new tree is strictly less than that of T which is not possible. Further, we have that T' is optimal.

Assume that we now have a tree

- i) which is optimal, and
- ii) in which the sibling property (P3) is true for the nodes at levels $l_{\max} - k$ to l_{\max} , for some k satisfying $0 \leq k < l_{\max} - 1$.

Change, if necessary, the assignment of children at level $l_{\max} - k - 1$ to the internal nodes at level $l_{\max} - k - 2$ so that the resulting tree is still D -ary, but now the sibling property (P3) is satisfied by the nodes at level $l_{\max} - k - 1$. A tree obtained after this re-assignment is clearly optimal. Moreover, the nodes at levels $l_{\max} - k$ to l_{\max} still satisfy the sibling property (P3) as the nodes at these levels which

were siblings before the rearrangement remain so after it. This process can be continued till we get a tree which is optimal and satisfies (P2) and (P3). This tree is a Huffman tree for P (Remark 1 and Theorem 1) and has L as its length sequence. \square

Even though the Huffman algorithm is restrictive in the range of optimal trees it constructs, Lemma 3 assures us that the algorithm can yield an optimal tree corresponding to any optimal length sequence.

Thus, for the case of $n = D^m$, any point of maximum has a Huffman tree with L_U as its length sequence. From Lemmas 1 and 2, the following theorem follows:

Theorem 3. *i) When n is a power of D , the points of maximum for the minimum expected length function \mathcal{L}_D are those $P = (p_1, p_2, \dots, p_n)$ from the set \mathcal{P}_n for which the sum of the lowest D probabilities is greater than or equal to its maximum probability, i.e.*

$$\sum_{i=n-D+1}^n p_i \geq p_1.$$

ii) When n is not a power of D , the uniform distribution U_n is the unique point of maximum for \mathcal{L}_D .

IV. CONCLUDING REMARKS

The points of maximum of the minimum expected length function have been completely characterized using a characterization of Huffman trees, and a chain of inequalities that Hwang used to show the Schur-concavity of the minimum expected length function. This result shows that the points of maximum known in the literature are all that exist.

REFERENCES

- [1] D. A. Huffman, “A Method for the Construction of Minimum-Redundancy Codes,” *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.
- [2] F. Hwang, “Generalized Huffman Trees,” *SIAM J. Appl. Math.*, vol. 37, no. 1, pp. 124–127, 1979.
- [3] N. Geckinli, “Two Corollaries to the Huffman Coding Procedure,” *IEEE Trans. Inf. Th.*, vol. 21, no. 3, pp. 342–344, 1975.
- [4] R. Gallager, “Variations on a Theme by Huffman,” *IEEE Trans. Inf. Th.*, vol. 24, no. 6, pp. 668–674, 1978.
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

Achieving Secrecy Capacity of Minimum Storage Regenerating Codes for all Feasible (n, k, d) Parameter Values

V. Arvind Rameshwar

Department of Electrical Communication Engineering
 Indian Institute of Science, Bengaluru
 Email: vrameshwar@iisc.ac.in

Navin Kashyap

Department of Electrical Communication Engineering
 Indian Institute of Science, Bengaluru
 Email: nkashyap@iisc.ac.in

Abstract—This paper addresses the problem of constructing secure exact-repair regenerating codes at the MSR point for all feasible values of the parameters. The setting involves a passive eavesdropper who is allowed to observe the stored contents of, and the downloads into, an l -subset of the n nodes of a distributed storage system (DSS). The objective is to achieve perfect secrecy between the eavesdropped symbols and the file stored on the DSS. Previous secure code constructions (most notably that by Rawat et al.) tackle the problem only for the restricted case wherein the number, d , of helper nodes aiding in the recovery of a failed node is equal to $n - 1$. This paper builds on Rawat’s work, by combining Gabidulin pre-coding and an MSR construction by Ye and Barg to prove the achievability of secrecy capacity at the MSR point for all allowed values of d .

I. INTRODUCTION

A distributed storage system (DSS) stores a file \mathbf{f} of size M (symbols over a finite field \mathbb{F}) on n storage nodes. The system possesses the “ k -out-of- n ” property, in that a data collector (DC) can recover the file by connecting to any k -subset of the nodes. The nodes, however, are prone to failure and the objective is to design schemes that allow for failed-node repair by contacting any d helper nodes, while preserving the “ k -out-of- n ” property. The work by Dimakis et al. [1] introduced the concept of *regenerating codes*, which address the problem of simultaneous repair and reconstruction while ensuring that each node stores no more than α independent symbols and each helper node passes on no more than β independent symbols to the failed node. Then from [1],

$$M \leq \sum_{i=1}^k \min\{\alpha, (d-i+1)\beta\}. \quad (1)$$

The upper bound describes a tradeoff between the parameters α and β , for a fixed M . Two extremal points of this trade-off curve are the minimum storage regeneration (MSR) and the minimum bandwidth regeneration (MBR) points. The MSR point, which is of interest to us, is where α is minimized for a given M . From [1] and the tradeoff curve (1), we have

$$(\alpha_{MSR}, \beta_{MSR}) = \left(\frac{M}{k}, \frac{M}{k(d-k+1)} \right).$$

Since MSR codes are equivalent to standard MDS array codes, the goal is to suitably augment MDS array code constructions with repair schemes. MSR codes that meet the capacity upper bound of (1) are described in [2], [6], [8] and [11]. In particular, Ye and Barg’s constructions in [11] allow for the parameter k to take on all feasible values (from 1 to n), and similarly, d to take any value in its permissible range of $k + 1$ to $n - 1$.

Now consider the *passive eavesdropper* setting, where an eavesdropper, Eve, is allowed to observe, over a long time, the stored contents of, and the downloads into, an l -subset of the n nodes. We need to ensure that Eve obtains no information about the file stored in the DSS.

Capacity upper bounds for *perfect secrecy* at the MSR and MBR points are provided in [5]. For the MSR point, work towards tightening the bound in [5] can be found in [3], [4] and [8]. While secure codes meeting the capacity upper bound at the MBR point in [5] have been constructed for all values of n, k, d [9], the task of constructing secure codes at the MSR point that achieve the improved capacity upper bound in [3] has been tackled only for the restricted case of $d = n - 1$.

In this work, we provide a secure code construction at the MSR point that achieves the capacity upper bound in [3] and [8] for all values of n, k, d , effectively closing the open problem of achieving secrecy capacity at the MSR point.

In Section II, we provide a formal description of the system model and discuss related literature. Section III describes the MSR code construction, and provides a proof of secrecy.

II. BACKGROUND AND RELATED WORK

In this section, we formally describe the system model, and provide details of Gabidulin-based pre-coding, and an overview of the MSR construction by Ye and Barg [11]. In what follows, the notation $[a : b]$ denotes the set of integers between a and b , both inclusive, i.e., $[a : b] = \{i \in \mathbb{Z} : a \leq i \leq b\}$. We use $[n]$ as shorthand for $[1 : n]$.

A. System Model

An (n, k, d) DSS consists of n storage nodes, indexed from 1 to n , that store in a distributed, coded fashion, the M

symbols (over a field \mathbb{F}) of a file \mathbf{f} . The symbols are drawn independently and uniformly at random from the field.

The recovery of the stored file follows the k -out-of- n property, i.e., it is sufficient to contact any subset of k nodes, to recover \mathbf{f} . Let \mathbf{c}_i , $i \in [n]$, denote the coded symbols stored in node i . Firstly, we require that each node stores no more than α independent symbols.

We assume that node failures in the system occur in stages, with no more than one failure at any stage. At stage t , we say that a node j is *active* if it does not fail in that stage. We operate in the *exact-repair* setting, wherein the downloads from d active helper nodes ($k+1 \leq d \leq n-1$) can exactly recover the contents of the failed node. In keeping with [1], our second constraint is that the failed node downloads no more than β independent symbols from any one helper node.

Now, suppose that node i has failed. Let $D_{j,i}$ denote the collection of random symbols sent by helper node j to i . If $H(X)$ represents the entropy of a random variable X , then

$$H(\mathbf{c}_i) \leq \alpha, \quad (2)$$

$$H(D_{j,i}) \leq \beta. \quad (3)$$

From [10], we know that exact-repair codes that satisfy (1) with equality must also satisfy (2) and (3) with equality.

Since at the MSR point, α is minimized for a given M , we have from (1) that $\alpha = M/k$. The minimum value of β then is $\beta = \alpha/(d-k+1)$. Hence,

$$(\alpha, \beta) = \left(\frac{M}{k}, \frac{M}{k(d-k+1)} \right).$$

Now consider the case where an eavesdropper, Eve, observes the downloaded symbols into an arbitrary l -subset \mathcal{E} of the nodes. We assume that each node in \mathcal{E} may fail multiple times, and in order to repair the same node over and over again, information is possibly downloaded from different sets of helper nodes. Thus, over time, Eve knows the stored contents of the nodes in \mathcal{E} , and has observed repair information for nodes in \mathcal{E} from all nodes not in \mathcal{E} . Let the random vector \mathbf{e} denote the symbols observed by Eve. Thus, \mathbf{e} consists of \mathbf{c}_i , $i \in \mathcal{E}$, as well as all the $D_{j,i}$, $i \in \mathcal{E}$, $j \notin \mathcal{E}$. If $\mathbf{f}^{(s)}$ ((s) for “secure”) is the file that we desire to store on the DSS, and $M^{(s)}$ is its size, the *perfect secrecy* condition then is: $I(\mathbf{f}^{(s)}; \mathbf{e}) = 0$, where $I(\mathbf{x}; \mathbf{y})$ is the mutual information between the random vectors \mathbf{x} and \mathbf{y} .

B. Related Work

The setting of the passive adversary was first discussed in [5], and an upper bound on the secrecy capacity for functional repair was derived to be

$$M^{(s)} \leq \sum_{i=l+1}^k \min\{\alpha, (d-i+1)\beta\},$$

where $l = |\mathcal{E}|$. Later work by Shah et al. [9] employed the Product-Matrix (PM) code construction to design a secure MSR coding scheme that achieved a maximum file size of $(k-l)(\alpha - l\beta)$. This was improved upon in [8] and [3], wherein the secrecy capacity was shown to be bounded as

$$M^{(s)} \leq (k-l)(1 - 1/(d-k+1))^l \alpha. \quad (4)$$

This upper bound was shown to be achievable in [8], for the case $n = d + 1$, using the concept of zigzag codes. Another achievability scheme, due to Rawat [7], uses a construction in Ye and Barg’s paper [11] to show the capacity upper bound in (4) being met, again when $n = d + 1$. In this paper, we build upon Rawat’s work to prove the achievability of the capacity upper bound in (4) for *all* feasible values of d , using an alternative construction from [11].

The tools we need are provided by the work of Huang et al. [4]. Recall that in a DSS, a given helper node j may in general belong to multiple repair groups (sets of helper nodes) for a given failed node i . A distributed storage code operating at the MSR point is said to be *stable* [4, Definition 7] if for each pair of nodes i and j , the information downloaded from node j to repair node i is the same across all repair groups for i containing the node j . In Lemma 7 of [4], it is shown that for DSSs based on a stable MSR code, the secrecy capacity of an $(n = d + 1, k, d)$ DSS is the same as that of an $(n > d + 1, k, d)$ DSS, when all other parameters are identical. This result, along with the observations of Rawat [7], in fact suffices to establish that the construction we describe in Section III achieves the secrecy capacity upper bound in (4). We, however, give a more direct proof, involving some ideas from hypergraph theory that may be of independent interest.

C. Preliminaries

Given a DSS that can store M symbols when $l = 0$, we augment our file of size $M^{(s)}$ with random symbols \mathbf{r} , where \mathbf{r} is a random vector of length $R = M - M^{(s)}$. Each random symbol in \mathbf{r} is drawn i.i.d. and uniformly at random from the field \mathbb{F} . We shall now describe the ingredients of our construction, namely the Gabidulin pre-coding procedure and the *d-optimal repair* MSR construction (for all parameters n, k, d), by Ye and Barg [11].

1) Gabidulin Pre-coding: Assume that we have an M -length vector $\mathbf{m} = (m_1, \dots, m_M)$, where each m_i , $1 \leq i \leq M$, is drawn from a finite field \mathbb{F} . Let \mathbb{B} be some sub-field of \mathbb{F} . Further, let points y_1, y_2, \dots, y_M , be elements of \mathbb{F} that are linearly independent over \mathbb{B} ($\dim_{\mathbb{B}}(\mathbb{F}) \geq M$).

The procedure for Gabidulin coding is:

- First, a linearized polynomial $p_{\mathbf{m}}(x)$ is constructed:

$$p_{\mathbf{m}}(x) = \sum_{i=0}^{M-1} m_{i+1} x^{|\mathbb{B}|^i}$$

- The polynomial is evaluated at the collection \mathcal{Y} of points y_1, y_2, \dots, y_M , yielding

$$p(\mathbf{m}, \mathcal{Y}) := (p_m(y_1), p_m(y_2), \dots, p_m(y_M)).$$

2) Ye and Barg construction: Here, we provide a brief description of the *d-optimal repair* construction.

Construction 1: First we shall introduce some notation: let $s = d - k + 1$ and let $\alpha = s^n$. Let \mathbb{F} be a finite field of size $|\mathbb{F}| \geq sn$ and let $\{e_a : a \in [0 : \alpha - 1]\}$ be the standard basis of \mathbb{F}^α over \mathbb{F} .

For an integer $a \in [0 : \alpha - 1]$, let $\mathbf{a} = (a_n, a_{n-1}, \dots, a_1)$ denote its s -ary representation so that $\mathbf{a} = \sum_{i=1}^n a_i s^{i-1}$. Suppose that $\{\lambda_{i,j}\}$ for $i \in [n]$ and $j \in [0 : s - 1]$ are s^n distinct elements in \mathbb{F} . We define the matrices A_i , $i \in [n]$, to be $\alpha \times \alpha$ diagonal matrices with the $(a, a)^{th}$ entry being λ_{i,a_i} . In other words,

$$A_i = \sum_{a=0}^{\alpha-1} \lambda_{i,a_i} e_a e_a^T. \quad (5)$$

We shall construct an $(n - k)\alpha \times n\alpha$ parity check matrix H for the MSR code \mathcal{C} as:

$$H = \begin{pmatrix} I & \dots & I & I \\ A_1 & \dots & A_{n-1} & A_n \\ A_1^2 & \dots & A_{n-1}^2 & A_n^2 \\ \vdots & \ddots & \vdots & \vdots \\ A_1^{n-k-1} & \dots & A_{n-1}^{n-k-1} & A_n^{n-k-1} \end{pmatrix} \quad (6)$$

where I is the $\alpha \times \alpha$ identity matrix.

In [11], the authors prove that the code \mathcal{C} obeys the “ k -out-of- n ” property, while storing exactly α independent symbols in each node. In addition, it is proved that the exact-repair requirement of the DSS is also met, ensuring that the contents of any one failed node can be exactly recovered from $\beta = \frac{\alpha}{d-k+1} = \frac{s^n}{s} = s^{n-1}$ symbols from each of d other active nodes.

We shall now describe the repair scheme. Let $\mathbf{c} = \{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ be a codeword of \mathcal{C} . Since $\alpha = s^n$, we shall index the symbols in \mathbf{c}_i by n -tuples from $[0 : s - 1]^n$. Let $\mathcal{S} = [0 : s - 1]$ and $\mathcal{S}^n = [0 : s - 1]^n$. The symbols in \mathbf{c}_i are indexed by the vectors $\mathbf{a} \in \mathcal{S}^n$, in lexicographic order, starting with the vector $\mathbf{0}$ and ending with the vector \mathbf{z} (the s -ary representation of $\alpha - 1$). In vectorized form, \mathbf{c}_i is the $\alpha \times 1$ column vector given as

$$\mathbf{c}_i = (c_{i,\mathbf{0}}, \dots, c_{i,\mathbf{z}})^T.$$

For a vector $\mathbf{a} \in \mathcal{S}^n$, let $(a_n, a_{n-1}, \dots, a_1)$ be its s -ary representation. Now, let $\mathbf{a}(i, u) \in \mathcal{S}^n$ be the vector obtained by substituting the symbol a_i in the s -ary representation of \mathbf{a} , with u , for $i \in [n]$ and $u \in [0 : s - 1]$. Thus,

$$\mathbf{a}(i, u) \equiv (a_n, a_{n-1}, \dots, a_{i+1}, u, a_{i-1}, \dots, a_1).$$

Assume that node i has failed and hence, \mathbf{c}_i needs to be recovered. Recall that each helper node j , $j \neq i$, sends exactly $\beta = s^{n-1}$ independent symbols to node i . For some $\mathbf{a} \in \mathcal{S}^n$, we define the set $\mathcal{S}_{\mathbf{a},i}^n$ as

$$\mathcal{S}_{\mathbf{a},i}^n := \{\mathbf{a}(i, u) : u \in \mathcal{S}\}.$$

Note that $|\mathcal{S}_{\mathbf{a},i}^n| = s$, for any $\mathbf{a} \in \mathcal{S}^n$. Furthermore,

$$\bigcup_{\mathbf{a}} \mathcal{S}_{\mathbf{a},i}^n = \mathcal{S}^n.$$

Thus, there exist $\beta = s^{n-1}$ distinct sets $\mathcal{S}_{\mathbf{a},i}^n$, the union of which is the entire set of s -ary n -tuples. We shall use these β distinct sets to index the symbols sent by node j to failed node i .

Now, let $D_{j,i}$ represent the symbols contributed by helper node j towards the repair of node i ($j \neq i$). Thus, from [11], $D_{j,i}$ is the row vector of the β symbols

$$\mu_{j,i}^{(\mathcal{S}_{\mathbf{a},i}^n)} = \sum_{u=0}^{s-1} c_{j,\mathbf{a}(i,u)} \quad (7)$$

for *distinct* sets $\mathcal{S}_{\mathbf{a},i}^n$, $\mathbf{a} \in \mathcal{S}^n$. Observe that \mathcal{C} is a stable MSR code, in the sense of [4, Definition 7].

III. SECURE MSR CODES FOR ALL PARAMETERS

In this section, we describe our construction of secure MSR codes for all feasible values of d , using arguments from [7].

Construction 2: Consider a file $\mathbf{f}^{(s)}$, that we intend storing on the DSS, of size

$$M^{(s)} = (k - l)(1 - 1/(d - k + 1))^l \alpha \quad (8)$$

symbols, over a field \mathbb{F} . The file size in (8) meets the secrecy capacity upper bound at the MSR point, derived in [3]. As in the Ye and Barg construction in Section II-C2, we take $\alpha = s^n$, where $s = d - k + 1$, for $k + 1 \leq d \leq n - 1$. We now describe our coding scheme:

- 1) **Gabidulin pre-coding:** To the information set of size $M^{(s)}$, we add $R = M - M^{(s)}$ random symbols (denoted by the vector \mathbf{r}), drawn i.i.d. and uniformly from the field \mathbb{F} , where $M = k\alpha$. Let this overall message $\mathbf{m} = (\mathbf{f}^{(s)}, \mathbf{r})$ be Gabidulin pre-coded by the procedure described in II-C1. Let

$$\mathbf{f} := p(\mathbf{m}, \mathcal{Y}) = (p_m(y_1), p_m(y_2), \dots, p_m(y_M)).$$

- 2) **Ye and Barg encoding:** Let H be the parity check matrix of the Ye and Barg code specified in Construction 1 of Section II-C2. The $k\alpha \times n\alpha$ generator matrix of the code, G , satisfies

$$GH^T = \mathbf{0},$$

where $\mathbf{0}$ denotes the $k\alpha \times (n - k)\alpha$ zero matrix. The code vector $\mathbf{c} = (\mathbf{c}_1, \dots, \mathbf{c}_n)$ to be stored in the nodes of the DSS is obtained as

$$\mathbf{c} = \mathbf{f}G \in \mathcal{C},$$

where \mathbf{f} is the pre-coded vector from Step 1. The i^{th} node of the DSS stores the vector \mathbf{c}_i of α symbols.

From the discussion in Section II-C2 and from [11], we know that the coding scheme described above is MSR and satisfies the exact-repair property for all values of d . The proof of secrecy follows next.

A. Proof of secrecy for $k + 1 \leq d \leq n - 1$

We shall follow the line of argument presented in [7]. Let the set of nodes that Eve eavesdrops on be $\mathcal{E} = \{i_1, i_2, \dots, i_l\}$. In the worst case, all nodes in \mathcal{E} have failed at least once. Note that, as before, we require $|\mathcal{E}| = l < k$. Further, let $D_{j,\mathcal{E}}$ represent the symbols sent by the j^{th} active

node ($j \in \mathcal{D} \subset [n] \setminus \mathcal{E}$, such that $|\mathcal{D}| = d$), for the repair of the nodes in \mathcal{E} . Hence,

$$D_{j,\mathcal{E}} = [D_{j,i_1} | D_{j,i_2} | \dots | D_{j,i_l}],$$

where the solid vertical lines represent concatenation.

Without loss of generality, we assume that $\mathcal{E} = \{n-l+1, n-l+2, \dots, n\}$, for, if otherwise, we can always reorder the nodes prior to the first node failure. To characterize the symbols downloaded by the nodes in \mathcal{E} , we make the following definition.

Definition III.1. (Symbol Matrix): A symbol matrix P corresponding to the repair scheme (7) is a 0-1 matrix of dimension $l\beta \times \alpha$ such that $D_{j,\mathcal{E}}^T = P\mathbf{c}_j$ for all $j \in [n] \setminus \mathcal{E}$.

In order to explicitly describe the entries of P , we require some notation. Recall from Section II that \mathcal{S}^n represents the set of vectors in $[0:s-1]^n$, and that $\mathcal{S}_{\mathbf{a},i}^n = \{\mathbf{a}(i,u) : u \in \mathcal{S}\}$, for $\mathbf{a} \in \mathcal{S}^n$. Now, define

$$\mathcal{S}_{i \leftarrow *}^n := \{(a_n, \dots, a_1) : a_i = *, a_j \in \mathcal{S} \text{ for } j \neq i\}.$$

Note that for any i , $|\mathcal{S}_{i \leftarrow *}^n| = s^{n-1}$.

For a vector $\mathbf{a} \in \mathcal{S}^n$ (or $\mathcal{S}_{i \leftarrow *}^n$ for some j), let $\mathbf{a}_{\setminus i}$ denote the vector obtained by puncturing \mathbf{a} in its i^{th} coordinate:

$$\mathbf{a}_{\setminus i} = (a_n, \dots, a_{i+1}, a_{i-1}, \dots, a_1).$$

Now, from the definition of the symbol matrix P , we have

$$P = \begin{bmatrix} P_n \\ \hline P_{n-1} \\ \hline \vdots \\ \hline P_{n-l+1} \end{bmatrix} \quad (9)$$

where each P_i , $i \in [n-l+1 : n]$ is a 0-1 matrix of dimensions $\beta \times \alpha$, such that $P_i \mathbf{c}_j = D_{j,i}^T = (\mu_{j,i}^{(\mathcal{S}_{\mathbf{a},i}^n)}, \dots, \mu_{j,i}^{(\mathcal{S}_{\mathbf{a},i}^n)})^T$, with $\mu_{j,i}^{(\mathcal{S}_{\mathbf{a},i}^n)}$ as in (7).

We now seek to characterize P_i , $i \in [n-l+1 : n]$, completely. Let the columns of P_i ($i \in [n-l+1 : n]$) be indexed by all the vectors in \mathcal{S}^n , listed in lexicographic order and let the rows of P_i be indexed by the vectors $\mathbf{b} \in \mathcal{S}_{i \leftarrow *}^n$, in lexicographic order.

From (7), we see that the row in P_i indexed by some $\mathbf{b} \in \mathcal{S}_{i \leftarrow *}^n$ contains exactly s 1's — these are in the columns indexed by the vectors $\mathbf{b}(i,u) = (b_n, \dots, b_{i+1}, u, b_{i-1}, \dots, b_1)$, for $u \in [0:s-1]$. All other entries of P_i are 0's. Note that the column indices containing a 1 entry differ in exactly their i^{th} coordinate. Explicitly,

$$[P_i]_{\mathbf{r},\mathbf{t}} = \begin{cases} 1, & \text{if } \mathbf{t}_{\setminus i} = \mathbf{r}_{\setminus i}, \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

for $\mathbf{r} \in \mathcal{S}_{i \leftarrow *}^n$ and $\mathbf{t} \in \mathcal{S}^n$.

Further, equation (9) coupled with equation (10) above, implies that each column of P contains exactly l 1's, one in each P_i .

Equations (9) and (10) completely characterize P . We add that $H(D_{j,\mathcal{E}}) = \text{rank}(P)$, since the symbols in \mathbf{c}_j are independent of one another.

As an example, consider the $(n, k, d, l) = (4, 2, 3, 1)$ DSS wherein Eve eavesdrops on the last (fourth) node. The symbol matrix P in this case is:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is easy to verify that $\text{rank}(P)$ above is 8, which in turn is equal to $H(D_{j,\mathcal{E}})$.

We intend to obtain a handle on the rank of P , in general. To this end, we claim that the following theorem holds true:

Theorem III.1. For $s = d - k + 1 \geq 2$,

$$H(D_{j,\mathcal{E}}) = s^{n-l}(s^l - (s-1)^l). \quad (11)$$

In other words, the rank of the symbol matrix P is $s^{n-l}(s^l - (s-1)^l)$.

We shall defer the proof of Theorem III.1 until later, and prove the following theorem, based on the validity of Theorem III.1.

Theorem III.2. The coding scheme of Construction 2 is secure, for $k+1 \leq d \leq n-1$, against a passive eavesdropper that has access to a set $\mathcal{E} \subset [n]$ of nodes, with $|\mathcal{E}| = l$.

Proof. The proof of the theorem is similar to the proof of Proposition 1 in [7]. We intend showing that $H(\mathbf{e}) \leq H(\mathbf{r})$ and $H(\mathbf{r}|\mathbf{e}, \mathbf{f}^{(s)}) = 0$, thereby implying (from the perfect secrecy lemma of [9]) that $I(\mathbf{f}^{(s)}; \mathbf{e}) = 0$. Let \mathcal{T} represent a group of $k-l$ nodes such that $\mathcal{T} \cap \mathcal{E} = \emptyset$. We know that

$$\mathbf{e} = (\mathbf{c}_i : i \in \mathcal{E}) \cup \left(\bigcup_{i \in \mathcal{E}} \bigcup_{j \in [n] \setminus \mathcal{E}} \{D_{j,i}\} \right).$$

Now, using the notation $\mathbf{c}_{\mathcal{E}} := (\mathbf{c}_i : i \in \mathcal{E})$, we have

$$\begin{aligned} H(\mathbf{e}) &= l\alpha + H\left(\bigcup_{i \in \mathcal{E}} \bigcup_{j \in [n] \setminus \mathcal{E}} \{D_{j,i}\} \mid \mathbf{c}_{\mathcal{E}}\right) \\ &= l\alpha + H\left(\bigcup_{i \in \mathcal{E}} \bigcup_{j \in \mathcal{T}} \{D_{j,i}\} \mid \mathbf{c}_{\mathcal{E}}\right) \\ &\leq l\alpha + H\left(\bigcup_{i \in \mathcal{E}} \bigcup_{j \in \mathcal{T}} \{D_{j,i}\}\right) \\ &\leq l\alpha + \sum_{j \in \mathcal{T}} H(D_{j,\mathcal{E}}) \\ &= ls^n + (k-l)(1 - (1-1/s)^l)s^n \\ &= ks^n - (k-l)(1-1/s)^l s^n \\ &= M - M^{(s)} = H(\mathbf{r}). \end{aligned} \quad (12)$$

The equality in (12) follows from the fact that

$$\begin{aligned} & H\left(\bigcup_{i \in \mathcal{E}} \bigcup_{j \in [n] \setminus (\mathcal{T} \cup \mathcal{E})} \{D_{j,i}\} \mid \mathbf{c}_{\mathcal{E}}, \bigcup_{i \in \mathcal{E}} \bigcup_{j \in \mathcal{T}} \{D_{j,i}\}\right) \\ & \leq \sum_{i \in \mathcal{E}} H\left(\bigcup_{j \in [n] \setminus (\mathcal{T} \cup \mathcal{E})} \{D_{j,i}\} \mid \mathbf{c}_{\mathcal{E}}, \bigcup_{j \in \mathcal{T}} \{D_{j,i}\}\right) \quad (13) \\ & = 0, \end{aligned}$$

the last equality holding since each summand in (13) equals 0 by the arguments used in the proof of Lemma 7 in [4].

Using the MDS array property of the Ye and Barg code and from Remark 8 of [8], it is possible to show that $H(\mathbf{r}|\mathbf{e}, \mathbf{f}^{(s)}) = 0$. We refer the reader to the proof of Proposition 1 in [7], for more details.

Now, from the perfect secrecy lemma in [9], the two conditions above imply that $I(\mathbf{f}^{(s)}; \mathbf{e}) = 0$, thereby proving that perfect secrecy holds. \square

We shall now proceed to the proof of Theorem III.1, beginning with the definitions of a few notions related to hypergraphs.

Definition III.2. (Incidence matrix) The incidence matrix (or vertex-edge incidence matrix) V of a hypergraph (X, E) is a 0-1 matrix of dimension $|V| \times |E|$, with the rows representing nodes and columns representing hyperedges, such that $V_{i,j} = 1$ if edge j is incident on vertex i , and 0 otherwise.

For a vector \mathbf{v} , we define its support to be the set of coordinates in which \mathbf{v} takes on non-zero values.

Definition III.3. (Connected hypergraph) A hypergraph (X, E) is said to be connected, if for every pair of nodes $(u, w) \in X \times X$, $u \neq w$, there exists an alternating sequence of nodes and hyperedges, $v_0, h_0, v_1, h_1, \dots, v_{m-1}, h_{m-1}, v_m$, ($m \in \mathbb{Z}_+$) with $v_0 = u$ and $v_m = w$, such that for $i \in [0 : m - 1]$, h_i is incident on both v_i and v_{i+1} . We call the sequence of hyperedges h_0, h_1, \dots, h_{m-1} as a *path* from u to w .

Now, we denote by $\mathcal{G}_{s,n}$, the n -dimensional regular hypergraph (X, E) with $|X| = s^n$ and $E \subset X^s$, with incidence matrix $V_{\mathcal{G}_{s,n}}$ defined as follows: let the rows of $V_{\mathcal{G}_{s,n}}$ be indexed by all the vectors in \mathcal{S}^n , listed in lexicographic order. Further, let the columns of $V_{\mathcal{G}_{s,n}}$ be indexed by the vectors $\mathbf{b} \in \mathcal{S}_{i \leftarrow *}^n$, (i ranging from n down to 1), where for any i , the vectors \mathbf{b} are listed in lexicographic fashion. Hence, the first s^{n-1} columns of $V_{\mathcal{G}_{s,n}}$ are indexed in lexicographic order by vectors in $\mathcal{S}_{n \leftarrow *}^n$, the next s^{n-1} columns are indexed by vectors in $\mathcal{S}_{(n-1) \leftarrow *}^n$, and so on. Thus, there are ns^{n-1} columns overall. The entries of $V_{\mathcal{G}_{s,n}}$ are

$$[V_{\mathcal{G}_{s,n}}]_{\mathbf{r}, \mathbf{t}} = \begin{cases} 1, & \text{if } \mathbf{r}_{\setminus i} = \mathbf{t}_{\setminus i} \text{ and } t_i = * \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

where $\mathbf{r} \in \mathcal{S}^n$ and $\mathbf{t} \in \mathcal{S}_{i \leftarrow *}^n$, $i \in [n]$.

The column in $V_{\mathcal{G}_{s,n}}$ indexed by the vector $\mathbf{b} \in \mathcal{S}_{i \leftarrow *}^n$ for some i , has exactly s 1's in precisely those rows \mathbf{t} for which $t_i = u$, $u \in [0 : s - 1]$ and $t_j = b_j$, for $j \neq i$. Moreover, each row of $V_{\mathcal{G}_{s,n}}$ has exactly n 1's.

With the aid of the definitions above, we wish to prove Theorem III.1. Two lemmas follow. The first (Lemma III.3) establishes that the transpose of the symbol matrix P can be thought of as the incidence matrix of a regular hypergraph having exactly s^{n-l} connected components. Our second lemma (Lemma III.4) proves that the rank of the incidence matrix corresponding to each of these connected components is $(s^l - (s-1)^l)$. We conclude by showing that the rank of P is precisely the sum of the ranks of the incidence matrices of these connected components.

Lemma III.3. P^T is the incidence matrix of a subgraph \mathcal{H} of $\mathcal{G}_{s,n}$. Further, the number of connected components in \mathcal{H} is s^{n-l} .

Proof. Recall that the symbol matrix P has a block matrix form, given in equation (9). Taking the transpose of each P_i , $i \in [n-l+1 : n]$ thus gives us:

$$[P_i^T]_{\mathbf{r}, \mathbf{t}} = \begin{cases} 1, & \text{if } \mathbf{t}_{\setminus i} = \mathbf{r}_{\setminus i}, \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

where $\mathbf{r} \in \mathcal{S}^n$ and $\mathbf{t} \in \mathcal{S}_{i \leftarrow *}^n$.

Since i ranges from $n-l+1$ to n only, by comparing the equation above with (14), we get that P^T is a submatrix of $V_{\mathcal{G}_{s,n}}$, containing only the first ls^{n-1} columns of $V_{\mathcal{G}_{s,n}}$. We denote by \mathcal{H} , the subgraph induced by this submatrix; thus, \mathcal{H} is a sub-hypergraph of $\mathcal{G}_{s,n}$.

Now, let t be the number of connected components in \mathcal{H} and let h_i , $1 \leq i \leq t$, be the number of hyperedges in each component \mathcal{H}_i . Then,

$$\sum_{i=1}^t h_i = l\beta = ls^{n-1}. \quad (16)$$

Pick some node indexed by vector \mathbf{v} so that \mathbf{v} belongs to connected component \mathcal{H}_j . Consider the collection of nodes \mathcal{W} , where $\mathcal{W} = \{\mathbf{w} \in \mathcal{S}^n : w_i = v_i, i \in [n-l] \text{ and } w_j \in [0 : s-1], j \in [n-l+1 : n]\}$. Note that the set \mathcal{W} includes the node \mathbf{v} . From equation (10), it is easy to verify that the sub-hypergraph of \mathcal{H} induced by the nodes in \mathcal{W} forms the connected component \mathcal{H}_j . This can be seen by choosing some node $\mathbf{x} \notin \mathcal{W}$. Node \mathbf{x} differs from any node in \mathcal{W} in at least one position $j \in [n-l]$. Hence, the row corresponding to \mathbf{x} in P^T will have as support, those columns where none of the nodes in \mathcal{W} have a 1 entry, thereby implying that there does not exist a path from \mathbf{x} to any of the nodes in \mathcal{W} . Observe that the number of nodes in \mathcal{H}_j is s^l .

Since the coordinates of nodes \mathbf{w} in \mathcal{H}_j in positions $i \in [n-l]$ are fixed, we puncture the vectors \mathbf{w} at these positions, to form the vectors \mathbf{w}' . Thus, the sets $\mathcal{S}_{i \leftarrow *}^n$, $i \in [n-l+1 : n]$ can be written as the sets $\mathcal{S}_{i \leftarrow *}^l$, $i \in [l]$, each set now containing vectors \mathbf{w}' .

The incidence matrix $V_{\mathcal{H}_j}$ has entries

$$[V_{\mathcal{H}_j}]_{\mathbf{r}, \mathbf{t}} = \begin{cases} 1, & \text{if } \mathbf{t}_{\setminus i} = \mathbf{r}_{\setminus i}, \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

where $\mathbf{r} \in \mathcal{S}^l$ and $\mathbf{t} \in \mathcal{S}_{i \leftarrow *}^l$, $i \in [l]$.

\mathcal{H}_j is precisely the hypergraph $\mathcal{G}_{s,l}$, having $h_j = ls^{l-1}$ edges. Since this is true for any $j \in [t]$, substituting in equation (16), we get that the number of connected components in \mathcal{H} equals s^{n-l} . \square

We shall now introduce an $(s-1)^l$ -dimensional code $\mathcal{C}^{\otimes l}$, of block length s^l , the parity check matrix of which will aid us in obtaining a handle on the rank of $V_{\mathcal{G}_{s,l}}$.

Consider the single parity check code \mathcal{C}_s of block length s over the field \mathbb{F} , having the $(s-1) \times s$ generator matrix G_s given by

$$G_s = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

The parity check matrix, H_s , of \mathcal{C}_s is simply the $1 \times s$ all-ones matrix. We then define the direct product code $\mathcal{C}_s^{\otimes l}$ (where \mathcal{C}_s is the underlying code), as the code generated by

$$G_s^{\otimes l} = \underbrace{G_s \otimes G_s \otimes \cdots \otimes G_s}_{l \text{ times}}$$

where \otimes denotes the Kronecker product. A codeword in $\mathcal{C}_s^{\otimes l}$ is of length s^l and can be described by an l -dimensional array. Each entry of the array (which is a coordinate of the codeword) is indexed by an s -ary l -tuple $\mathbf{v} = (v_1, v_2, \dots, v_l) \in \mathcal{S}^l$.

The code $\mathcal{C}_s^{\otimes l}$ has the property that each array element \mathbf{v} is involved in l parity check equations, one along each coordinate $i \in [l]$. In other words, for every symbol $\mathbf{v} \in \mathcal{S}^l$, there exists a parity check equation indexed by a vector $\mathbf{b} \in S_{i \leftarrow *}^l$, $i \in [l]$, such that $\mathbf{b}_{\setminus i} = \mathbf{v}_{\setminus i}$. Moreover, the parity check equation along coordinate j is the sum of those codeword symbols that differ from \mathbf{v} in only their j^{th} coordinate.

Formally, the code $\mathcal{C}_s^{\otimes l}$ can be described by the $ls^{l-1} \times s^l$ parity check matrix H having entries

$$H_{\mathbf{r}, \mathbf{t}} = \begin{cases} 1, & \text{if } \mathbf{t}_{\setminus i} = \mathbf{r}_{\setminus i}, \\ 0, & \text{otherwise} \end{cases} \quad (18)$$

where $\mathbf{r} \in S_{i \leftarrow *}^l$, $i \in [l]$ and $\mathbf{t} \in \mathcal{S}^l$. Thus, each row of H corresponds to a parity check equation \mathbf{b} , that finds the sum of symbols $\mathbf{v} \in \mathcal{S}^l$, which differ only in that coordinate of the l -tuple, i , in which $b_i = *$.

Lemma III.4. *The parity check matrix H of $\mathcal{C}_s^{\otimes l}$ is equal to $V_{\mathcal{G}_{s,l}}^T$. Further, $\text{rank}(V_{\mathcal{G}_{s,l}}^T) = s^l - (s-1)^l$.*

Proof. The first part of the lemma is obvious from equations (14) and (18). We observe that $\text{rank}(G_s^{\otimes l}) = \prod_{i=1}^l \text{rank}(G_s) = \text{rank}(G_s)^l = (s-1)^l$. Thus, the rank of the parity check matrix H is equal to $s^l - (s-1)^l$. \square

Using Lemmas III.3 and III.4, we shall now prove Theorem III.1.

Proof. Recall from Lemma III.3 that for any node $\mathbf{v} \in \mathcal{S}^n$ of the hypergraph \mathcal{H} , the connected component \mathcal{H}_j containing

\mathbf{v} consists of nodes in the set $\mathcal{W} = \{\mathbf{w} \in \mathcal{S}^n : w_i = v_i, i \in [n-l] \text{ and } w_j \in [0 : s-1], j \in [n-l+1 : n]\}$.

It is now possible to permute the rows \mathbf{v} of P^T in lexicographic order of $\mathbf{v}' = (v_{n-l}, \dots, v_1)$ so that all the rows $\mathbf{v} \in \mathcal{S}^n$ corresponding to a fixed value of \mathbf{v}' occur together. Thus, the first s^l rows of P^T are indexed by vectors \mathbf{v} such that $\mathbf{v}' = (0, 0, \dots, 0)$, the next s^l rows are indexed by vectors \mathbf{v} with $\mathbf{v}' = (0, 0, \dots, 1)$ and so on. Each collection of s^l rows corresponding to a particular value of $\mathbf{v} = (v_{n-l}, \dots, v_1)$ forms the incidence matrix of a connected component.

From (17), we observe that the supports of the rows corresponding to any two connected components \mathcal{H}_i and \mathcal{H}_j , $i \neq j$, are disjoint. Hence, the rank of P^T is equal to the sum of the ranks of the incidence matrices of the connected components of hypergraph \mathcal{H} , induced by P^T . Since each connected component is precisely the hypergraph $\mathcal{G}_{s,l}$ (from the proof of Lemma III.3), we get that

$$\text{rank}(P^T) = s^{n-l}(\text{rank}(V_{\mathcal{G}_{s,l}})) = s^{n-l}(s^l - (s-1)^l),$$

where the first equality follows from Lemma III.3 and the second from Lemma III.4. \square

ACKNOWLEDGEMENTS

N. Kashyap's work was supported by a Swarnajayanti Fellowship awarded by the Department of Science and Technology, Government of India.

REFERENCES

- [1] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran. Network coding for distributed storage systems. *IEEE Transactions on Information Theory*, 56(9):4539–4551, Sept 2010.
- [2] S. Goparaju, A. Fazeli, and A. Vardy. Minimum storage regenerating codes for all parameters. *IEEE Transactions on Information Theory*, 63(10):6318–6328, Oct 2017.
- [3] S. Goparaju, S. El Rouayheb, A. R. Calderbank, and H. V. Poor. Data secrecy in distributed storage systems under exact repair. *CoRR*, abs/1304.3156, 2013.
- [4] K. Huang, U. Parampalli, and M. Xian. On secrecy capacity of minimum storage regenerating codes. *IEEE Transactions on Information Theory*, 63(3):1510–1524, March 2017.
- [5] S. Pawar, S. El Rouayheb, and K. Ramchandran. On secure distributed data storage under repair dynamics. In *2010 IEEE International Symposium on Information Theory*, pages 2543–2547, June 2010.
- [6] K. V. Rashmi, N. B. Shah, and P. V. Kumar. Optimal exact-regenerating codes for distributed storage at the msr and mbr points via a product-matrix construction. *IEEE Transactions on Information Theory*, 57(8):5227–5239, Aug 2011.
- [7] A. S. Rawat. Secrecy capacity of minimum storage regenerating codes. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 1406–1410, June 2017.
- [8] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath. Optimal locally repairable and secure codes for distributed storage systems. *IEEE Transactions on Information Theory*, 60(1):212–236, Jan 2014.
- [9] N. B. Shah, K. V. Rashmi, and P. V. Kumar. Information-theoretically secure regenerating codes for distributed storage. In *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, pages 1–5, Dec 2011.
- [10] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran. Distributed storage codes with repair-by-transfer and nonachievability of interior points on the storage-bandwidth tradeoff. *IEEE Transactions on Information Theory*, 58(3):1837–1852, March 2012.
- [11] M. Ye and A. Barg. Explicit constructions of high-rate mds array codes with optimal repair bandwidth. *IEEE Transactions on Information Theory*, 63(4):2001–2014, April 2017.

Data-pooling and multi-task learning for enhanced performance of speech recognition systems in multiple low resourced languages

Madhavaraj A and A G Ramakrishnan
MILE Lab, Electrical Engineering, Indian Institute of Science, Bangalore 560012
madhavaraja@iisc.ac.in, agr@iisc.ac.in

Abstract—We present two approaches to improve the performance of automatic speech recognition (ASR) systems for Gujarati, Tamil and Telugu. In the first approach using data-pooling with phone mapping (DP-PM), a deep neural network (DNN) is trained to predict the senones for the target language; then we use the feature vectors and their alignments from other source languages to map the phones from the source to the target language. The lexicons of the source languages are then modified using this phone mapping and an ASR system for the target language is trained using both the target and the modified source data. This DP-PM approach gives relative improvements in word error rates (WER) of 5.1% for Gujarati, 3.1% for Tamil and 3.4% for Telugu, over the corresponding baseline figures. In the second approach using multi-task DNN (MT-DNN) modeling, we use feature vectors from all the languages and train a DNN with three output layers, each predicting the senones of one of the languages. Objective functions of the output layers are modified such that during training, only those DNN layers responsible for predicting the senones of a language are updated, if the feature vector belongs to that language. This MT-DNN approach achieves relative improvements in WER of 5.7%, 3.3% and 5.2% for Gujarati, Tamil and Telugu, respectively.

Index Terms: Multi-task learning, data-pooling, deep neural networks, phone mapping, alignments, senone posteriors, cross-lingual training, multilingual training, parameter sharing, speech recognition, Gujarati, Tamil, Telugu.

I. INTRODUCTION

Building a large vocabulary, continuous speech recognition system requires a huge corpus of transcribed speech so as to effectively estimate the acoustic model parameters, and a huge corpus of text in order to estimate the language model parameters. Although such corpora exist for English and a few other languages, there are many languages for which corpora are not readily available, and collecting such data is a cumbersome and time-consuming task. For such low-resourced languages, the traditional way of acoustic modeling results in a high word error rate. Assuming there exists similarity in acoustic units across languages, it is possible to use data from a high-resourced language in order to estimate acoustic models for a low-resourced target language [1]. In this work, we focus only on the acoustic modeling of the target language by borrowing transcribed speech corpora from one or more source languages.

Lal and King [2] have used tandem features in a cross-lingual training setting, where a neural network is trained across several languages to predict articulatory features and the outputs from this neural network are used as features to train the hidden Markov model (HMM) based acoustic models. Lu et al. [3] have used subspace Gaussian mixture models (SGMM) to learn global parameters from multiple languages and the state-specific parameters are learned from the target language data. They have also experimented maximum *a posteriori* adaptation

to reduce the mismatch between the source and the target languages' SGMM global parameters. They have also extended SGMM-based cross-lingual training with l_1 -regularization for estimating the state vectors [4]. This is shown to provide less word error rate, while also overcoming the problem of numerical instability. Schultz and Weibel [5] have built a language-independent speech recognition system by combining acoustic models from multiple source languages using language-separate, language-mixed and language-tagged combining methods. Manohar et al. [6] have used phone-cluster adaptive training to obtain the acoustic model parameters by linear combination of a canonical Gaussian mixture model. The mean vectors of the Gaussian mixture models for each state are parameterized by a state-vector, which is estimated through the procedure proposed by Gales [7].

Miao et al. [8] show the advantage of using deep maxout networks (DMN) in acoustic modeling. DMNs, which possess the property of dropout, have shown very good performance, particularly for low-resource languages. Sahraeian and Compernolle [9] have used manifold learning technique to derive a non-linear feature transformation from filter-bank space to articulatory space. They have used intrinsic spectral analysis and deep neural networks (DNNs) to convert acoustic features to articulatory features and used them in cross and multi-lingual training settings. Mohan and Rose [10] have used multi-task deep neural networks along with low-rank matrix factorization of the weight matrices for multi-lingual speech recognition systems. They have obtained a reduction of 44% in the number of parameters without compromising much on the word error rate (WER), when the DNNs are trained only on one hour of target language data. Heigold et al. [11] present an empirical comparison on mono-, multi- and cross-lingual training of deep neural networks for eleven languages with a total data of 10k hours in a distributed manner. They have also shown that performing multilingual training on top of cross-lingual training gives an additional relative reduction of 5% in the WER. Data pooling of closely related languages [1] has resulted in improvements in the performance of automatic speech recognition (ASR) systems for under-resourced languages. They have shown that having two hours of data from a closely related non-target language is equivalent to having one hour of target language data. In the semi-supervised acoustic model (AM) refinement technique used by [12], an initial model is trained using a small amount of transcribed corpus and the AM is refined iteratively by choosing the unlabeled speech and its decoded utterance (decoded by the AM) based on some confidence measures.

In the work reported here, we propose two approaches to improve the performance of automatic speech recognition systems for Gujarati, Tamil and Telugu. The motivation to utilize the speech data from all the three languages to build ASR for any one of the languages arises from the similarity in the phonology of Indian languages [13].

85% of the phones are common among Tamil, Telugu and Gujarati. Hence, for building an ASR for one of these languages as the target language, we can leverage the acoustic information from the other two languages also, for better modeling of the senone distributions. Towards this purpose, we have considered two distinct approaches, wherein the acoustic information is captured at the level of the (i) data (data pooling with phone mapping) or (ii) model (multi-task DNN).

The rest of this paper is organized as follows: In section 2, we describe our first approach of training and using a DNN to automatically map phones from source language(s) to a target language and then pooling all the source and target data together to build the speech recognition system. Section 3 describes the procedure to train a multi-task DNN using data from all the languages to predict the senones of all the languages and how the objective function is changed such that the weights are updated in a specific manner so that the first few layers capture the common acoustic information across all the languages. In section 4, we discuss the baseline system and the systems developed based on the approaches described in sections 2 and 3, and provide their results. Finally, in section 5, we conclude the paper and indicate a possible future research direction for this problem.

II. DATA POOLING WITH PHONE MAPPING

In the literature, data pooling has been used with an universal phone set [14], for cross-lingual training. However, in our approach of data pooling with phone mapping (DP-PM), we map the phones of the source languages to those of the target language using a deep neural network, trained only on the target language data. We then use this map to modify the phonetic transcriptions of the source languages to suit the target language and train the speech recognition system by all the data pooled together and fine-tune the DNN for the target language. To our knowledge, data pooling has not been used in this manner earlier. The steps involved are illustrated below. In all our experiments, the lexicon was designed by us by incorporating the pronunciation rules for the languages [15], [16]. For more details, refer [17],[18].

A. DNN training for the target language

We have used the standard procedure for training as given in s5 WSJ Kaldi recipe [19]. First we extract 39-dimensional features from the target data, consisting of mel-frequency cepstral coefficients (MFCC), delta and delta-delta features, and train a monophone model for each phone in the target language (with a total of 1000 Gaussian densities) for 40 iterations. Using the alignments from the monophone model, we then build *tri1* models, which are triphone, context-dependent HMM models with a total of 2500 states and 15000 Gaussian densities. From the alignments of *tri1* model, we then train *tri2* HMM model, which is based on a combination of linear-discriminant analysis (LDA) and maximum likelihood linear transformation (MLLT). From the alignments from *tri2*, we train *tri3*, a speaker-adaptive model (LDA-MLLT-SAT) and finally we align the data using *tri3* model. Using the probability density function (pdf) indices (also referred to as senones) from these alignments as desired target labels, and spliced MFCC features (with 5 each of left- and right-contexts) as input features, we train a *tridnn* model which is a 7-layer DNN with 2048 hidden sigmoid activations in each layer. The weights are randomly initialized and trained for 15 epochs. The learning rate is set as 0.008 for the first 4 epochs and is halved for

each subsequent epoch. This DNN is now able to predict the posterior probability of a HMM state's pdf, given any input feature vector.

B. Generating alignments for the source languages

We now train the *tri3* models for the source languages independently using the procedure explained in the previous section and then with respect to this *tri3* model, we align the source data (also referred to as source alignments).

It is to be noted that the HMMs for the source and target languages are trained with their respective phone-sets. In order to pool all the data together, we need the data from all the languages to have a common phonetic transcription with respect to the target language's phone-set. We explain in the next section as to how we use the DNN to convert the phonetic transcription from any source language to a particular target language.

C. Mapping of phones from the source to target language

We propose this approach based on the assumption that acoustic similarities exist across languages and the function that maps such a similarity can be extracted in a data-driven fashion. We use the DNN, trained as explained in section II.A, to map the phones from any source language to the target language. Let x_s and y_s be a feature vector and its corresponding senone-id of the source language. Let $g_s(\cdot)$, $g_t(\cdot)$ be the functions that map senones to phones for the source and the target languages, respectively, and $f(\cdot)$ be the non-linear function representing the prediction of the DNN.

We pass all the feature vectors of the source language to the DNN and predict the senone values (denoted by $f(x_s)$). From these predictions and the senone-to-phone mapping functions, we calculate the conditional density $P(g_t(f(x_s))|g_s(y_s))$. This density function P can be calculated by counting (and then normalizing) the number of times the feature vectors belonging to a phone in the source language are recognized by the DNN as any other phone in the target language. The final function $m(\cdot)$, which maps a source language phone (ϕ_s) to a target language phone (ϕ_t^*) is given by,

$$m(\phi_s) = \phi_t^* = \arg \max_{\phi_t} P(\phi_t|\phi_s)$$

Using the function $m(\cdot)$, we can modify the lexicon and the phonetic transcription of any source language so that it has only the phones from the phone set of the target language.

D. Data-pooling and training

Once the source data is mapped to the format required by the target language, we pool the utterances from the source as well as the target languages and train an ASR system starting from *mono* to *tridnn* as described in section II.A.

E. Fine-tuning the DNN for the target language

At the end of the previous step, we get a DNN trained with the pooled data. Now, we fine-tune this DNN by training with only the target language data with a learning rate of 0.0008 for 5 epochs. This fine-tuned DNN can now be used as the acoustic model for decoding the test utterances of the target language.

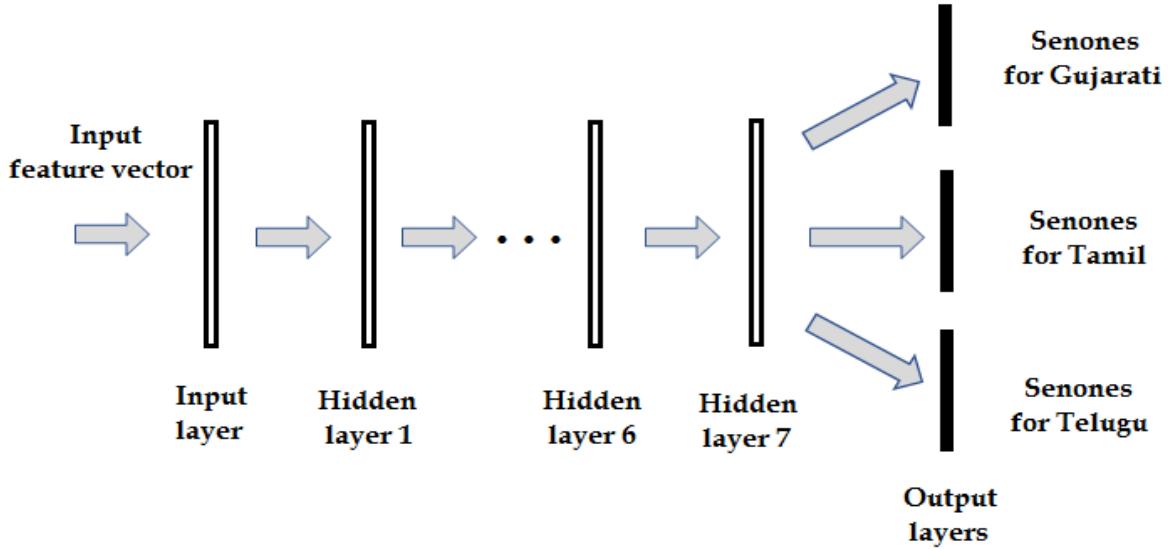


Fig. 1. Multi-task deep neural network architecture to predict senones for the three languages for any given input vector. The shared hidden layers learn feature representations common to all the languages and the three output layers perform the classification for the individual languages.

III. MULTI-TASK DNN APPROACH

In our second approach, we have used a multi-task deep neural network (MT-DNN) with multiple output layers (one for each language), as shown in Fig. 1. This architecture is similar to the one in [20]. The procedure involved in training and using such an MT-DNN as the acoustic model is described below.

A. Generating alignments for the source and the target languages

We use the procedure illustrated in section II.A and build the system from *mono* to *tri3* and using the *tri3* models, we generate the alignments for each language independently. The number of senones may differ from language to language and thus the senones do not have a straightforward correspondence between any two languages. We learn the senone correspondence through the above MT-DNN architecture, by modifying the training procedure as explained in the subsections below.

B. Features and targets for training the MT-DNN

Let x be a feature vector and y , its corresponding senone target. The feature vector can come from any one of, say, L languages. Each training example to the MT-DNN should be of the form $\{x, [y_1, y_2, \dots, y_L]\}$, where y_l is the desired target (in one-hot vector encoding format) for the l^{th} output layer. The i^{th} entry of the vector y_l is defined as,

$$y_l^i = \mathbb{1}(x \in l)\mathbb{1}(y = i) \quad \forall 1 \leq l \leq L$$

where $\mathbb{1}(\cdot)$ is the indicator function.

C. Modified loss function for training the MT-DNN

Normally, a feature vector belonging to a particular language is assigned zero as the desired target for all the other languages [11]. However, in the context of MT-DNN, since acoustic similarities exist

across the languages, it is inappropriate to force the MT-DNN to predict zeros as senone-posteriors for the other languages. We handle this issue by modifying the loss function in such a way that we update only those layers responsible for predicting the senones for the language to which the feature vector belongs. This modified loss function for the l^{th} layer \hat{L}^l is given by,

$$\hat{L}^l(z_l, y_l) = L(z_l, y_l)\mathbb{1}(x \in l)$$

where z_l is the actual predicted vector, y_l is the desired target vector at the l^{th} layer and $L(\cdot)$ is cross-entropy loss function. We now train the MT-DNN using this modified loss function with the feature vectors (in the format specified in section 3.2) from all the languages. We have used Keras [21] library to train this MT-DNN for 15 epochs and ported the trained network back to Kaldi format for the fine-tuning process. The learning rate was fixed at 0.008 for the first 4 epochs and reduced by half for the subsequent epochs. The main reason to train such an architecture is to ensure that the layers 1 through 7 learn feature representations that are common to all the languages and at the same time, increase the discriminability of every output layer.

Only the layers up to layer 7 learn the common representation, whereas the output layers do not learn any common representation. In other words, when a feature vector comes from a particular language, only the output layer corresponding to that language is updated. In order for the output layers to benefit from this training method, we can set the desired targets for the non-target output layers to be a predefined posterior vector instead of zeros. In such a case, there is no need to modify the loss function for training and all the output layers can be allowed to update for feature vectors from any of the languages. We hypothesize that this training procedure will further increase the performance of MT-DNN, which is yet to be experimented.

D. Fine-tuning the MT-DNN for the target language

Once the MT-DNN is trained, we retain only those layers of the MT-DNN that predict the output for the target language desired and remove the rest of the layers, thus having only one output layer. Now, we fine-tune this network for 5 epochs by using data from only the desired target language with a learning rate of 0.0008. This network can then be used as the acoustic model for decoding.

The advantage of using MT-DNN over DP-PM method is that there is no need to train the entire model set from *mono* to *tri3* once again. It is sufficient to fine-tune the DNN only for the desired target language and use it directly for testing.

IV. EXPERIMENTS AND RESULTS

All our experiments have been conducted on the transcribed speech corpus given by Microsoft [22]. The training data consists of transcribed speech corpus of 40 hours for training, 5 hours for validation and 4.2 hours for testing, for each of the three languages, namely Gujarati, Tamil and Telugu. We have created the trigram language models using only the training data's text corpora. The CMU Indic frontend lexicon provided for each language has been used as the pronunciation dictionary. The acoustic models for the baseline systems have been built as per the procedure explained in section II.A.

Based independently on (i) data-pooling with phone mapping and (ii) multi-task DNN approaches, we have built two systems as per the procedures illustrated in sections 2 and 3, respectively. Table 1 compares the word error rates of the acoustic model (AM) of the baseline DNNs with respect to the AMs of the DNNs obtained by the two proposed methods.

Table 1 reveals that for the validation datasets, the DP-PM method gives relative improvements in WER of 1.3% for Gujarati, 1.6% for Tamil and 2.3% for Telugu. On the other hand, MT-DNN provides the best relative improvements of 3.9% for Gujarati, 1.7% for Tamil and 4.1% for Telugu.

The same trend can be seen for the blind test data as well. The DP-PM model achieves relative improvements of 5.1%, 3.1% and 3.4% over the baseline in the word error rates for Gujarati, Tamil and Telugu, respectively. The MT-DNN model results in a marginal improvement and the relative improvements achieved over the baseline are 5.7%, 3.3% and 5.2%, respectively.

Thus, our best performing MT-DNN based method gives the lowest WERs of 24.3%, 32.0% and 30.2% on the blind test data for Gujarati, Tamil and Telugu languages, respectively. We can further reduce the error rates on the test data by using both the training and the validation datasets for building the acoustic and language models.

V. CONCLUSION AND FUTURE WORK

We have followed two approaches, namely data-pooling with phone mapping and multi-task DNN for cross-lingual training of the ASR for Gujarati, Tamil and Telugu languages. The first approach pools the data together by mapping the phones from the source languages to the target language, which gives relative improvements of 5.1%, 3.1% and 3.4% in the WERs for Gujarati, Tamil and Telugu test datasets, respectively. The second approach involves learning DNN model parameters from the pooled data using multi-task learning technique

with a modified loss function. This achieves relative improvements in the WERs of 5.7%, 3.3% and 5.2% for Gujarati, Tamil and Telugu, respectively.

Our future work will involve extending the multi-task learning approach by using mean statistic of the senone-posterior outputs for the feature vectors belonging to a particular senone class as desired targets for the MT-DNN, instead of predicting zeros for the non-target languages, without modifying the loss function.

REFERENCES

- [1] C. Van Heerden, N. Kleynhans, E. Barnard, and M. Davel, "Pooling ASR data for closely related languages," in *SLTU 2010: Proc. 2nd Workshop on Spoken Languages Technologies for Under-resourced languages*, 2010, pp. 17–23.
- [2] P. Lal and S. King, "Cross-lingual automatic speech recognition using tandem features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2506–2515, 2013.
- [3] L. Lu, A. Ghoshal, and S. Renals, "Cross-lingual subspace gaussian mixture models for low-resource speech recognition," *IEEE/ACM Trans. Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 17–27, 2014.
- [4] ———, "Regularized subspace gaussian mixture models for cross-lingual speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on*, 2011, pp. 365–370.
- [5] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.
- [6] V. Manohar, C. B. Srinivas, and S. Umesh, "Acoustic modeling using transform-based phone-cluster adaptive training," in *Automatic Speech Recognition and Understanding (ASRU), IEEE Workshop on*, 2013, pp. 49–54.
- [7] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.
- [8] Y. Miao, F. Metze, and S. Rawat, "Deep maxout networks for low-resource speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec 2013, pp. 398–403.
- [9] R. Sahraeian and D. V. Compernolle, "Crosslingual and multilingual speech recognition based on the speech manifold," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2301–2312, 2017.
- [10] A. Mohan and R. Rose, "Multi-lingual speech recognition with low-rank multi-task deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conf. on*, 2015, pp. 4994–4998.
- [11] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conf. on*, 2013, pp. 8619–8623.
- [12] M. Chellappiyadharshini, A. Toffy, R. Srinivasa, and M. V. Rama-subramanian, "Semi-supervised and active-learning scenarios: Efficient acoustic model refinement for low resource indian language," in *19th Annual Conference of International Speech Communication Association (Interspeech 2018)*, 2018.
- [13] K. V. Vijay Girish, V. Veena, and A. G. Ramakrishnan, "Relationship between spoken Indian languages by clustering of long distance bigram features of speech," in *India Conference (INDICON), 2016 IEEE Annual*. IEEE, 2016, pp. 1–6.
- [14] J. L. Hieronymus, "ASCII phonetic symbols for the world's languages: Worldbet," 1994.

TABLE I

COMPARISON OF WERs FOR THE BASELINE, DATA-POOLING WITH PHONE MAPPING (DP-PM) AND MULTI-TASK DNN (MT-DNN) MODELS ON VALIDATION AND TEST SETS. RELATIVE IMPROVEMENT IN WER (IN %) WITH RESPECT TO THE BASELINE IS GIVEN IN PARENTHESES FOR EACH CASE.

Method	Gujarati		Tamil		Telugu	
	Val. set	Test set	Val. set	Test set	Val. set	Test set
Baseline	18.8 (NA)	25.7 (NA)	32.8 (NA)	33.1 (NA)	30.6 (NA)	31.9 (NA)
DP-PM	18.6 (1.3)	24.4 (5.1)	32.3 (1.6)	32.1 (3.1)	29.9 (2.3)	30.8 (3.4)
MT-DNN	18.1 (3.9)	24.3 (5.7)	31.3 (1.7)	32.0 (3.3)	29.3 (4.1)	30.2 (5.2)

- [15] A. G. Ramakrishnan and M. Laxmi Narayana, “Grapheme to phoneme conversion for Tamil speech synthesis,” in *Proc. Workshop in Image and Signal Processing (WISP-2007), IIT Guwahati*, 2007, pp. 96–99.
- [16] A. G. Ramakrishnan, R. D. Sequiera, S. S. Rao, and H. R. Shiva Kumar, “Transliteration of Indic languages to Kannada with a user-friendly interface,” in *Advance Computing Conference (IACC), 2015 IEEE International*. IEEE, 2015, pp. 998–1001.
- [17] A. Madhavaraj and A. G. Ramakrishnan, “Design and development of a large vocabulary, continuous speech recognition system for Tamil,” in *2017 14th IEEE India Council International Conference (INDICON)*. IEEE, 2017, pp. 1–5.
- [18] A. Madhavaraj, H. R. Shiva Kumar, and A. G. Ramakrishnan, “Online speech translation system for Tamil,” in *19th Annual Conference of International Speech Communication Association (Interspeech 2018)*, 2018.
- [19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.
- [20] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conf. on*, 2013, pp. 7304–7308.
- [21] F. Chollet, “Keras,” <https://github.com/fchollet/keras>, 2015.
- [22] *Data provided by SpeechOcean.com and Microsoft*. Microsoft, 2018.

On the Role of Linear, Mel and Inverse-Mel Filterbank in the Context of Automatic Speech Recognition

Hemant K. Kathania
Dept. of ECE, NIT Sikkim
India
hemant.ece@nitsikkim.ac.in

Syed Shahnawazuddin
Dept. of ECE, NIT Patna
India
s.syed@nitp.ac.in

Waquar Ahmad
Dept. of ECE, NIT Calicut
India
waquar@nitsikkim.ac.in

Nagaraj Adiga
Dept. of CS, University of Crete
Greece
nagaraj@csd.uoc.gr

Abstract—In the context of automatic speech recognition (ASR), the power spectrum is generally warped to the Mel-scale during front-end speech parameterization. This is motivated by the fact that, human perception of sound is nonlinear. The Mel-filterbank provide better resolution for low-frequency contents while a greater degree of averaging happens in the high-frequency range. The work presented in this paper aims at studying the role of linear, Mel and inverse-Mel filterbanks in the context of speech recognition. It is well known that, when speech data is from high-pitched speakers like children, there is a significant amount of relevant information in the high-frequency region. Hence, down-sampling the information in that range through Mel-filterbank reduces the recognition performance. On the other hand, employing inverse-Mel or linear-filterbanks are expected to be more effective in such cases. The same has been experimentally validated in this work. To do so, an ASR system is developed on adults' speech and tested using data from adult as well as child speakers. Significantly improved recognition rates are noted for children's as well adult females' speech when linear or inverse-Mel filterbank is used. The use of linear filters results in a relative improvement of 21% over the baseline.

Index Terms: Filterbank, speech recognition, VTLN, pitch scaling.

I. INTRODUCTION

The task of developing a automatic speech recognition (ASR) system can be broken down into three major components namely, front-end speech parameterization, training acoustic models and language models. In this paper, our focus is on the first component i.e., front-end speech parameterization. The basic motivation for front-end speech parameterization is to derive a compact representation for the raw speech waveform after discarding the irrelevant informations. In the context of ASR, the speaker and environment dependent acoustic attributes need to be suppressed. Consequently, the ASR system becomes speaker and ambiance independent. In addition to that, since the raw data is represented in a compact manner, the overall complexity of training the system parameters as well as network search and decoding is reduced significantly.

A number of techniques have been proposed over the years for extracting front-end features from raw speech data. Among

those, the one based on Mel-frequency cepstral coefficients (MFCC) [1] has been the dominant one. During MFCC feature extraction process, the magnitude/power spectra of the short-time frame of speech under analysis is warped to the Mel-scale using a filterbank. Mel-scale warping is motivated by the findings of psychoacoustics that suggest that human perception of different frequency components is nonlinear. In other words, the use of Mel-filterbank is to mimic the human perception mechanism. The Mel-filterbank provide better resolution for low-frequency contents while a greater degree of averaging happens in the high-frequency range. As a result, the spectral information present in the high-frequency region of speech is down-sampled by Mel-scale warping. Since the speaker characteristics are predominantly reflected in the high-frequency components, the use of Mel-filterbank was observed to degrade the performance of automatic speaker recognition system [2] [3] [4] [5]. Motivated by this fact, the use of linear-filterbank was proposed in that work. On the other hand, Mel-scale warping helps in the case of ASR since the speaker characteristics get suppressed.

In this paper, we have studied the relative importance of linear and Mel-filterbanks in the context of ASR. For the sake of completeness, the role of inverse-Mel-filterbank is also explored. This study is motivated by the fact that, in the case of children's speech, a significant amount of relevant spectral information is present in the higher-frequency region. Therefore, wideband speech data (sampled at 16 kHz rate) is preferred in the case of children's ASR. As mentioned earlier, the resolution of Mel-filterbank decreases as the frequency is increased. Hence, down-sampling the spectra is not beneficial in those cases where the speech data is from high-pitched child speakers. This is also true in the case of adult females as observed in this work. On the contrary, providing equal resolution to all the frequencies or resolving higher-frequency contents with greater discrimination should be the preferred choice. In other words, using linear or inverse-Mel-filterbank will be more effective.

In order to verify our claims, separate set of front-end features were extracted by applying Mel-, inverse-Mel- and linear-filterbanks. Next, using each type of features, separate

ASR systems were trained on adults' speech data from both male and female speakers. The ASR system was evaluated using two different test sets. The first test set consisted of speech data from adult male and female speakers while the second one was comprised of the data from children. To get better insight, the adults' speech test set was further split into two parts based on the gender of the speaker. The use of Mel-filterbank was noted to be more effective when the test speech data was from adult male speakers. In those cases when speech data was from adult females or children, employing linear or inverse-Mel-filterbank was observed to be more effective. In order to further boost confidence in those observations, linear frequency warping through vocal-tract length normalization (VTLN) [6] and explicit pitch scaling was also explored to suppress the ill-effects induced by other speaker-dependent acoustic mismatch factors. Even after the inclusion of VTLN or pitch scaling, the use linear-filterbank was noted to be more effective when the test data was from high-pitch speakers (adult females and children).

The rest of this paper is organized as follows: In Section II, motivation for studying the role of filterbanks in ASR is discussed. In Section III, the experimental evaluations demonstrating the effectiveness of linear and inverse-Mel-filterbanks are presented. Finally, the paper is concluded in Section IV.

II. MOTIVATION FOR STUDYING THE ROLE OF FILTERBANKS IN ASR

As mentioned earlier, the MFCC features are one of the most dominant front-end features in the context of automatic speech recognition [7]. Given the raw speech data, the steps involved in the extraction of MFCC features are as follows: Speech signal is first processed through a pre-emphasis filter in order to emphasize the higher-frequency components. Next, short-time frames of speech are created using overlapping Hamming windows. Typically, the duration of the analysis window is 20-30 ms with an overlap of 50%. This is followed by deriving the frequency domain representation for each of the short-time frames. Discrete Fourier transform (DFT) is used for this purpose. The phase information is discarded from the resulting short-term spectrum. The magnitude or the power spectra is then warped to Mel-scale using a set non-linearly spaced filters. The Mel-filterbank is a set triangular Mel-weighted filters. Next, logarithmic compression is performed followed by the application of discrete cosine transform (DCT) to derive a set of de-correlated cepstral coefficients. Finally, a low-time filtering operation is performed to discard the higher-order coefficients. Only the first 13 coefficients are retained and those are collectively known as MFCC features.

The primary idea behind warping the linear frequencies to Mel-scale is to mimic the nonlinear behavior of human perception mechanism. The frequency resolution of the Mel-filterbank decrease as one moves towards the high-frequency region. This fact is evident from the spectral plot shown in Figure 1 (top pane). The short-time log-compressed power spectrum corresponding to the central portion of a voiced frame of speech is plotted. The 40-channel Mel-filterbank is

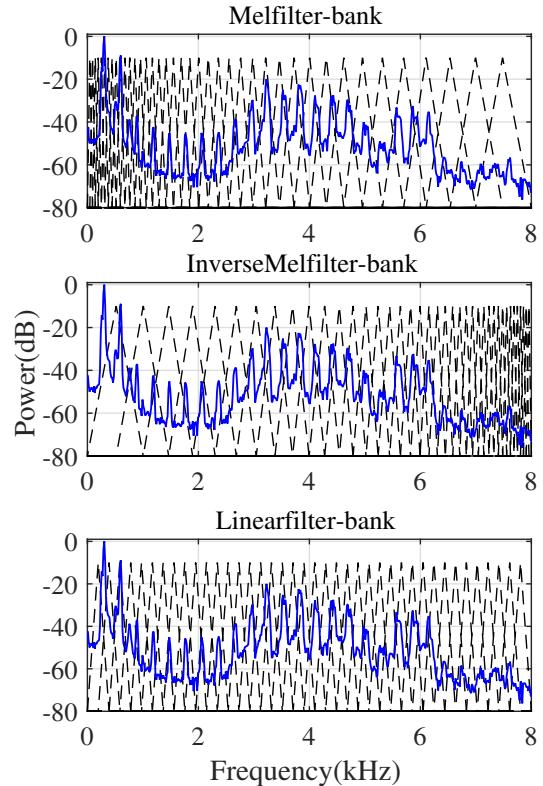


Fig. 1: Log-compressed power spectrum corresponding to the central portion of a voiced frame of speech from high-pitched child speaker. The 40-channel Mel-, inverse-Mel- and linear filterbanks are superimposed over the spectrum.

superimposed over the spectrum. The speech data used for this analysis is from a high-pitched child speaker. Further, wide-band speech data is used for all the analyses presented in this paper. As clearly visible from the plots, the degree of averaging is more in the high-frequency region. This behavior of Mel-filterbank has an added advantage that the speaker-dependent acoustic attributes are smoothed out. This, in turn, is beneficial for ASR task where speaker independence is highly desired.

When dealing with children's speech or speech from high-pitched speakers like adult females, the down-sampling of spectral information in high-frequency components has a downside. As already stated, there is a significant amount of spectral information in the higher-frequency region that is important for ASR. Earlier works have shown that, the formant frequencies are scaled up in the case of children's speech [8]–[10]. To demonstrate this characteristic of speech, the log-compressed power spectra corresponding to the central portion of vowel /IY/ are plotted in Figure 2. Scaling up of formant frequencies in the case speech data from adult female and child are easily noticeable. At the same time, the power is significantly high even in the 4-8 kHz frequency range. On the other hand, the power in 7-8 kHz frequency range is very less when the data is from adult male speaker. The spectral information in the high-frequency region should also be effectively preserved to improve the recognition performance with

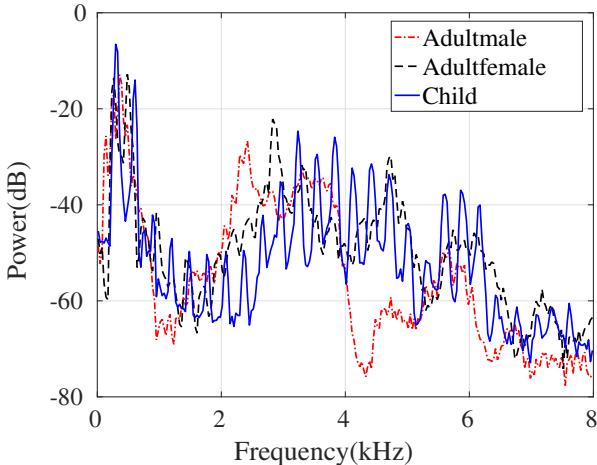


Fig. 2: Power spectra for vowel /IY/ extracted from the speech data belonging to adult male, adult female and children speakers, respectively. A short-time frame corresponding to the central portion of the vowel was used for this analysis.

respect to high-pitched speakers.

Motivated by these observations and findings of earlier works on children's speech, the role of inverse-Mel and linear-filterbanks are studied in this paper. The inverse-Mel-scale is defined as the complement of Mel-scale [11]. Unlike Mel-filterbank, better resolution is obtained in the higher-frequency region. This is evident from the log-compressed power spectrum with inverse-Mel-filterbank superimposed over it which is shown in Figure 1 (middle pane). The front-end features obtained by replacing the Mel-filterbank with inverse-Mel-filterbank are referred to as inverse-MFCC (IMFCC) in the remainder of this paper. The linear-filterbank provides equal resolution to all the frequency components and the same is evident from Figure 1 (bottom pane). The front-end features obtained by using linear-filterbank are called linear-frequency cepstral coefficients (LFCC) in this work. In the following section, we present the experimental evaluations demonstrating the relative effectiveness of MFCC, IMFCC and LFCC features in the context of ASR.

III. EXPERIMENTAL EVALUATIONS

The simulation studies performed for evaluating the relative effectiveness of MFCC, IMFCC and LFCC features are presented in this section.

A. Experimental setup

1) *Speech corpora:* The speech data used for training the ASR system was obtained from the **British English** speech corpus WSJCAM0 [12]. The train set created from WSJCAM0 consisted of 15.5 hours of speech data from 92 male/female adult speakers. The total number of utterances in the train set is 7,852 with a total 132,778 words. In order to evaluate the effectiveness of the explored front-end features, three different test sets were created. The details of those test sets are as follows:

- **AD-Set:** This test set was derived from the WSJCAM0 corpus and consisted of 0.6 hours of speech from 20 adult

male as well as female of speakers with a total of 5,608 words.

- **ADF-Set:** This test set was derived by splitting AD-Set and consisted of nearly 0.3 hours of speech from 10 adult female speakers with a total of 2,864 words.
- **CH-Set:** For evaluating recognition performance with respect to children's speech, a test set derived from PF-STAR [13] **British English speech database** was employed. This test set consisted of 1.1 hours of speech data from 60 child speakers with a total of 5,067 words. The age of the child speakers in this test set was in between 4 – 14 years.

The experimental studies reported in this paper were performed on wide-band (WB) speech data (sampled at 16 kHz rate). The PF-STAR database is originally sampled at 22,050 samples per second, so down-sampling was done for consistency.

2) *Front-end feature extraction:* In order to extract the three kinds of front-end features, speech data was first high-pass filtered with pre-emphasis factor being 0.97. Short-time frames were then created using overlapping Hamming windows of length 20 ms with frame-shift of 10 ms. For MFCC, IMFCC as well as LFCC, 40-channel filterbank was used to extract the 13-dimensional base features. Next, the base features were time-spliced by appending 4 frames to the left and to the right of the current analysis frame to it. The resulting 117-dimensional features vectors were then projected to 40 dimensional space using linear discriminant analysis (LDA) and maximum-likelihood linear transformation (MLLT) to derive the final feature vectors. This was followed by application of cepstral mean and variance normalization (CMVN) to all the front-end feature kinds. In addition to CMVN, feature normalization was done using feature-space maximum likelihood linear regression (fMLLR) for boosting robustness towards speaker-dependent variations [14].

3) *ASR system architecture:* The ASR systems were developed on the 15.5 hours adults' speech data from the WSJCAM0 speech corpus. The Kaldi speech recognition toolkit [15] was used for ASR system development and evaluation. Context-dependent hidden Markov models (HMM) were used in this work. Decision tree-based state tying was performed to fix the maximum number of tied-states (senones) at 2000. Observations densities for the HMM states were modeled using deep neural networks (DNN) [16], [17]. The fMLLR-normalized feature vectors were time-spliced once more considering a context size of 9 frames. The number of hidden layers in the DNN was chosen as 8. Each of the hidden layers consisted of 1024 hidden nodes with *tanh* non-linearity. The initial learning rate was selected as 0.015 which was reduced to 0.002 in 20 epochs. Extra 10 epochs were employed after reducing the learning rate. The minibatch size for neural net training was selected as 512. The initial state-level alignments employed in DNN training were generated using a Gaussian-mixture-model-based system.

While decoding adults' speech test, the standard MIT-Lincoln 5k Wall Street Journal bigram language model (LM)

TABLE I: WERs for the adults' and children's speech test sets with respect to the acoustic models trained on adults' speech. The WERs are given for the cases when MFCC, IMFCC and LFCC features are used to train DNN-HMM-based systems.

Test set	WER (in %)			Relative imp. over MFCC with LFCC (%)	
	Acoustic Feature				
	MFCC	IMFCC	LFCC		
AD-Set	5.87	5.93	6.11	- 4.1	
CH-Set	19.37	18.14	16.35	15.6	

was used. The MIT-Lincoln LM has a perplexity of 95.3 with respect to adults' test set with no out-of-vocabulary (OOV) words. The employed lexicon consisted of 5,850 words along with the pronunciation variations. While decoding the children's speech test, on the other hand, a 1.5k bigram LM was used. This bigram LM was trained on the transcripts of speech data in PF-STAR after excluding the test set. A lexicon consisting of 1,969 words including the pronunciation variations was used. The word error rate (WER) metric was used for evaluating the recognition performance.

B. Evaluation results

The baseline WERs for the adults' and children's speech test sets obtained by using MFCC features are presented in Table I. On comparing the WERs for AD-Set and CH-Set, a huge difference is noted. One of the factors for the observed difference is that the Mel-scale warping leads to down-sampling of spectral information in high-frequency components as discussed earlier. Consequently, the use of IMFCC and LFCC improves the recognition performance with respect to children's speech as given the WERs enlisted in Table I. At the same time, the recognition rates for adults' speech are noted degrade when LFCC features are used. But the loss incurred in the case of adults' speech is much less when compared to the gain obtained for children's speech. This fact is highlighted by the percentage relative improvement over MFCC obtained by using LFCC features given in the last column of Table I.

It may be argued that, by including children's speech data in the train set, the differences will be reduced as reported in earlier works. In order to demonstrate that by folding sufficient amount of speech data into training the ill-effects of down-sampling the spectral information present in high-frequency region cannot be addressed, the adult test set was split into two parts based on the gender of the speaker. The test set created by taking speech data only from female speakers (ADF-Set) was then decoded using the adult data trained acoustic models. The WERs for this study are given in Table II. The use of IMFCC and LFCC features is noted to reduce the WER significantly when compared to MFCC features. It is to note that the training set derived from WSJCAM0 database contains a sufficient amount of speech data from adult female speakers [12]. Despite that, providing higher resolution to

high-frequency components (IMFCC) or equal resolution to all the frequency components (LFCC) helps.

It may also be argued that, there several other factors of acoustic mismatch that lead to degradation in the recognition performance especially in the case of children's speech. In order to counter it, we have explored the role of VTLN and pitch scaling to reduce the acoustic mismatch resulting from higher fundamental and formant frequencies noted in the case of children's speech. These studies are presented in the following section.

C. Application of VTLN and pitch scaling

It is well known that, the vocal organs of children and adult females are smaller when compared to that of adult males [8]–[10], [18]–[20]. As a consequence, formant frequencies are upscaled when the speech data is either from adult females or children. Linear-frequency warping through vocal-tract length normalization (VTLN) [6] is reported to address the ill-effects of formant scaling [21], [22]. VTLN was implemented by extracting acoustic features after changing the linear frequency warping factor. The warp factor value was varied from 0.88 to 1.12 in steps of 0.02. The warped feature vectors were then forced-aligned against the acoustic model under the constraints of the first-pass hypothesis. The first-pass hypothesis was, in turn, obtained by decoding the unwarped features. The set of features that resulted in highest likelihood were chosen to be optimal. The optimally warped feature vectors were then re-decoded after performing fMLLR-based feature normalization. The effect of concatenating VTLN and fMLLR on the MFCC, IMFCC, and LFCC features are demonstrated using WERs given in Table II. Large reductions in WER are noted by the application of VTLN in the case of children's speech. The observed reductions in the case of adult females is not that large. At the same time, the LFCC features are still noted to be superior to MFCC for both ADF-Set and CH-Set test sets.

Apart from formant scaling, even the fundamental frequency or pitcg is noted to change due to differences in vocal-tract geometry. Consequently, the fundamental frequency is observed to be higher in the case of children as well as adult female speakers. Pitch-induced acoustic mismatch severely degrades the ASR performance as reported in [23], [24]. The ill-effects of pitch variations can be compensated by explicit pitch modification as reported in [25]. Motivated by that work, we have also explored pitch modification in order to improve the recognition performance with respect to high-pitched speakers. The pitch scaling technique reported in [26] was explored for this purpose. The tunable pitch compensation factor (semitone) was varied from 12 to 12 in steps of 1 to vary the pitch of the speech data being analyzed. The optimal compensation factor was chosen via a maximum-likelihood grid described earlier.

The WERs obtained by suitably modifying the pitch are given in Table II. Similar to the case of VTLN, large reductions in WER are observed in the case of children's speech. Even for adult females, the reductions in WER are significant. The use of LFCC features is noted to be superior in this case as

TABLE II: WERs for adult females' and children's speech test sets with respect to acoustic models trained on adults' speech. The WERs are given for the cases when MFCC, IMFCC and LMFCC features are used to train the DNN-HMM-based ASR systems. The WERs are also tabulated for the cases when VTLN and explicit pitch scaling (PS) are employed for reducing the acoustic mismatch.

Acoustic feature	Test set	WER (in %)		
		Baseline	VTLN	PS
MFCC	ADF-Set	6.35	6.11	5.67
	CH-Set	19.37	17.00	13.11
IMFCC	ADF-Set	6.10	5.93	5.28
	CH-Set	18.14	16.56	12.86
LFCC	ADF-Set	5.94	5.84	5.23
	CH-Set	16.35	14.89	12.19

well. It is to note that, the reduction in WER is larger in the case of pitch scaling than that obtained with the application VTLN. Pitch scaling is performed by re-sampling the speech data followed by time-scale modification. Re-sampling the speech data results in rescaling of the formant frequencies as well. Consequently, VTLN is done in a implicit manner when explicit pitch modification is performed.

IV. CONCLUSION

In this paper, we have studied the role Mel-, inverse-Mel- and linear-filterbanks in the context of ASR task. The presented work is motivated by the fact that there is a significant amount relevant spectral information present in the high-frequency region when the speech data is from adult female and child speakers. Consequently, down-sampling the spectral information in that range through Mel-filterbank reduces the recognition performance. The inverse-Mel and linear-filterbanks provide better resolution to the high-frequency components. Therefore, significant improvements are noted when IMFCC or LFCC features are used when the speech data being transcribed is from adult female or child speakers. In order to further boost our confidence in the observed improvements, the role of VTLN and explicit pitch scaling also explored. Even after the application of VTLN or pitch scaling, LFCC features are noted to be better than MFCC features.

REFERENCES

- [1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, August 1980.
- [2] M. Russell, S. D'Arcy, and L. Qun, "The effects of bandwidth reduction on human and computer recognition of children's speech," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1044–1046, Dec 2007.
- [3] S. D'Arcy and M. Russell, "A comparison of human and computer recognition accuracy for childrens speech," in *Proc. INTERSPEECH*, 2005, pp. 2187–2200.
- [4] M. R. Qun Li, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc. ICSLP2002*, Sept. 2002.
- [5] D. B. Serdar Yildirim, Shrikanth Narayanan and S. Khurana, "Acoustic analysis of preschool childrens speech," in *Proc. 15th ICPhS Barcelona*, 2003, pp. 949–952.
- [6] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, January 1998.
- [7] R. Vergin, D. O'Shaughnessy, and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-idenpendent continuous-speech recognition," *IEEE Trans. On ASSP*, vol. 7, no. 5, pp. 525–532, Sept. 1999.
- [8] R. D. Kent, "Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies," *JHSR*, vol. 9, pp. 421–447, 1976.
- [9] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, September 2007.
- [10] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.
- [11] Z. J. Yegnanarayana B., Prasanna S.R.M. and G. C. S., "Combining evidence from source, suprasegmental and spectral features for a fixed-text speaker verification system," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 4, pp. 575–582, July 2005.
- [12] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, May 1995, pp. 81–84.
- [13] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.
- [14] S. P. Rath, D. Povey, K. Vesely, and J. Černocký, "Improved feature processing for deep neural networks," in *Proc. INTERSPEECH*, 2013.
- [15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.
- [16] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.
- [17] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 20, no. 1, pp. 30–42, 2012.
- [18] S. Eguchi and I. J. Hirsh, "Development of speech sounds in children." *Acta oto-laryngologica. Supplementum*, vol. 257, pp. 1–51, 1969.
- [19] A. Potamianos and S. Narayanan, "Robust Recognition of Children Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, November 2003.
- [20] S. Lee, A. Potamianos, and S. S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, March 1999.
- [21] S. Ghai, "Addressing Pitch Mismatch for Children's Automatic Speech Recognition," Ph.D. dissertation, Department of EEE, Indian Institute of Technology Guwahati, India, October 2011.
- [22] R. Serizel and D. Giuliani, "Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children," *Natural Language Engineering*, April 2016.
- [23] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. INTERSPEECH*, 2016.
- [24] R. Sinha and S. Shahnawazuddin, "Assessment of pitch-adaptive front-end signal processing for children's speech recognition," *Computer Speech & Language*, 2017.
- [25] H. K. Kathania, W. Ahmad, S. Shahnawazuddin, and A. B. Samaddar, "Explicit pitch mapping for improved children's speech recognition," *Circuits, Systems, and Signal Processing*, September 2017.
- [26] W. Ahmad, S. Shahnawazuddin, H. K. Kathania, G. Pradhan, and A. B. Samaddar, "Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion," in *INTERNSPEECH*, 2017.

Speaking-Rate Adaptation of Automatic Speech Recognition System through Fuzzy Classification based Time-Scale Modification

S. Shahnawazuddin
Dept. of ECE
NIT Patna
s.syed@nitp.ac.in

Hemant K. Kathania
Dept. of ECE NIT Sikkim
India
hemant.ece@nitsikkim.ac.in

Waquar Ahmad
Dept. of ECE NIT Calicut
India
waquar@nitc.ac.in

Nagaraj Adiga
Dept. of CS University of Crete
Greece
nagaraj@csd.uoc.gr

B. Tarun Sai
Dept. of ECE NIT Patna
India
s.syed@nitp.ac.in

Abstract—In this paper, we study the role of speaking-rate adaptation (SRA) of automatic speech recognition (ASR) systems. The performance of an ASR system is reported to degrade when the speaking-rate is either too fast or too slow. In order to simulate such a situation, an ASR system was trained on adults' speech and used for transcribing speech data from adult as well as child speakers. Earlier studies have shown that, speaking-rate is significantly lower in the case of children when compared to adults. Consequently, the recognition performance for children's speech was noted to be very poor in contrast to adults' speech. To improve the recognition performance with respect to children's speech, speaking-rate was explicitly changed using time-scale modification (TSM). A recently proposed TSM approach based on fuzzy classification of spectral bins has been explored in this regard. The fuzzy-classification-based TSM technique is reported to be superior to state-of-the-art approaches. Effectiveness of the said TSM technique has not been studied yet in the context of ASR. The experimental studies presented in this paper show that SRA based on fuzzy classification results in a relative improvement of 30% over the baseline.

Index Terms—Speaking-rate adaptation, automatic speech recognition, time-scale modification, fuzzy classification.

I. INTRODUCTION

The task of modifying the duration of a signal without changing the frequency contents is referred to as time-scale modification (TSM). Several different techniques for TSM have been proposed over the years [1]. Changing the duration of a signal is beneficial for numerous applications. Common examples are modifying the length of a music signal in order to fit within a prescribed time-slot and viewing a video in slow-motion. Another important area of signal processing where TSM finds application is the task of changing the speaking-rate (SR). Different people speak at different rates, some being slow while others being fast. Consequently, it is difficult to recognize the spoken words when the speaking-rate is extremely fast. The same problem is faced by automatic speech recognition (ASR) systems as well. Even though an ASR system is expected to be insensitive towards variations in

speaking-rate, earlier studies have shown that the recognition performance degrades significantly when the speaking-rate is too fast or too slow [2]–[5]. Hence, TSM can be used to enhance the robustness of ASR systems towards speaking-rate variations.

In this paper, we have experimentally studied the ill-effects of speaking-rate variation on the recognition performance of an ASR system. To simulate such a task, an ASR system was developed using speech data from adult speakers. Next, the developed system was used for transcribing speech data from adult as well as child speakers. The task of decoding children's speech using acoustic models trained on adults' data is an example where the differences in the speaking-rate are highly pronounced. In the case of children, the average phoneme duration is longer [4], [6], [7]. Hence, the speaking-rate for children is lower than that for adults. In addition to that, variability in speaking-rate is higher among the child speakers themselves. Consequently, the error rates were noted to be much higher in the case of children's speech when compared to task of decoding adults' data.

In order to normalize the differences in speaking-rate, we have studied the role of TSM in this work. In other words, to improve the recognition performance with respect to children's speech, TSM has been employed for suitably increasing the speaking-rate for child speakers. The approach for TSM explored in this study is the one based on fuzzy classification of spectral bins [8]. The fuzzy-classification-based TSM technique has been proposed recently and is reported to be better than the existing similar approaches. Its effectiveness in the context of ASR system has not been studied yet. Speaking-rate adaptation (SRA) via fuzzy-classification-based TSM is noted to be highly effective as demonstrated by the experimental studies presented in this paper. Even though TSM was explored in some of the earlier reported works [9], [10], acoustic modeling based on Gaussian mixture models (GMM) was employed in those works. In this paper, we have used acoustic

modeling approaches based deep neural networks (DNN) [11] and long short-term-memory (LSTM) [12] recurrent neural networks (RNN) for experimental evaluations.

The rest of this paper is organized as follows: In Section II, the proposed speaking-rate adaptation technique is described. In Section III, the experimental evaluations demonstrating the effectiveness of the proposed approach are presented. Finally, the paper is concluded in Section IV.

II. SPEAKING-RATE ADAPTATION

A. Motivation

As stated earlier, speaking-rate for adult and child speakers differ significantly. It was observed in [13]–[15] that, production as well as perception of phones get affected by the variation in speaking-rate. Therefore, when speaking-rate is exceptionally fast or slow, the recognition performance of an ASR system is observed to degrade severely. A typical example of such a scenario is the task of decoding speech data from children using an ASR system trained on adults' speech. The average vowel durations in the case of children's speech are longer than those for the adults. An increase in average vowel duration implies that the speaking-rate for children is lesser than that for adults [4].

In order to gain a better understanding about the aforementioned differences, speaking-rate was computed for adults and children using a large number of continuous speech utterances from both the groups of speakers. The speech data used for this purpose was derived from two separate **British English** corpora, namely WSJCAM0 [16] (adults' speech) and PESTAR [17] (children's speech). Further details about those speech databases are summarized later in Section III. In this work, we have quantified the speaking-rate in terms of number of phonemes per second. The variation of speaking-rate in the case of adult and child speakers is demonstrated using the histograms shown in Figure 1. We have used 500 utterances from each group of speakers in order to derived those histograms. It is evident from Figure 1 that, the mean speaking-rate for adult speakers is almost two times greater than that for the children. Since data driven machine learning techniques are commonly used for the task of speech recognition, extreme acoustic mismatch occurs when an ASR system trained on adults' speech is used for transcribing children's data. In order to improve the recognition performance, speaking-rate normalization through TSM was explored in some of the earlier reported works on children's ASR [9], [10]. Pitch-synchronous overlap and add (PSOLA) algorithm was used for SRA in [9]. On the other hand, pitch-synchronous time-scaling [18] was studied in [10]. Motivated by those works, we have also explored SRA for improving children's speech recognition in this paper. In this regard, a recently proposed TSM approach based on fuzzy classification of spectral bins is explored. The fuzzy-classification-based technique for TSM is reported to be superior to state-of-the-art approaches.

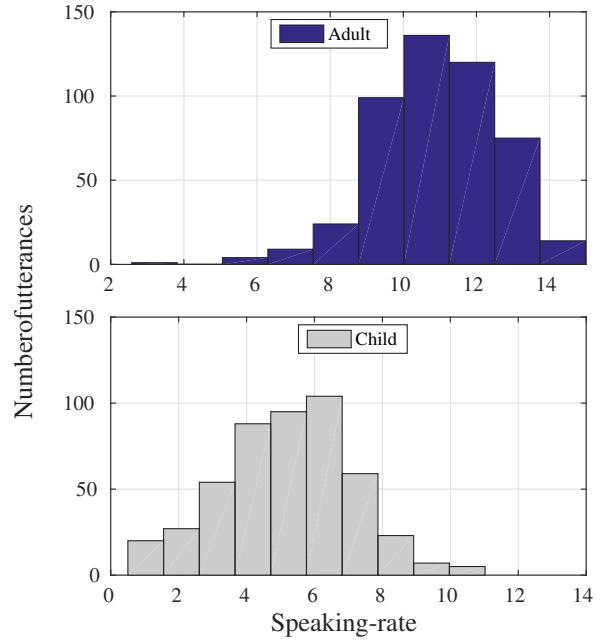


Fig. 1: Histograms depicting the variation in speaking-rate for adult and child speakers quantified in terms of number of phonemes per second. Speaking-rate was computed using 500 utterances from adult and children.

B. Fuzzy classification based time-scale modification

Any audio signal can be considered to consist of three different components, viz. sinusoidal, noise, and transient [19]. In the context of TSM, the most challenging part is to preserve the quality of those three components simultaneously. To deal with this challenge, some of the earlier reported TSM approaches resorted to a binary classification of the spectral bins. However, binary classification has a serious drawback since the energy in each spectral bin is actually a combination of energy from each of those three aforementioned components [8]. To overcome this limitation, the spectral bins should belong to each of the three classes at the same time with an associated degree of class-membership. In other words, the approach for classifying the spectral bins should be *fuzzy* [20] instead of being binary. Motivated by this fact, the characteristics of an audio signal were quantified using fuzzy classification in a recently proposed technique for TSM [8]. When compared to the approach based on harmonic-percussive separation [21], fuzzy-classification-based TSM was observed to be better. Motivated by its success, we have explored the effectiveness of this technique in the context of automatic speech recognition.

C. Effect of TSM on speaking-rate

The effect of TSM on the duration of given speech signal is shown through a set of time domain waveforms in Figure 2. In order to increase the speaking, the given speech signal was compressed by a factor of 0.7. On the other, a scaling factor of 1.4 was used for decreasing the speaking-rate. The corresponding spectrograms are also shown in Figure 2. From the figure, it is evident that the shape of the speech signal as well as the formant transitions are preserved even after

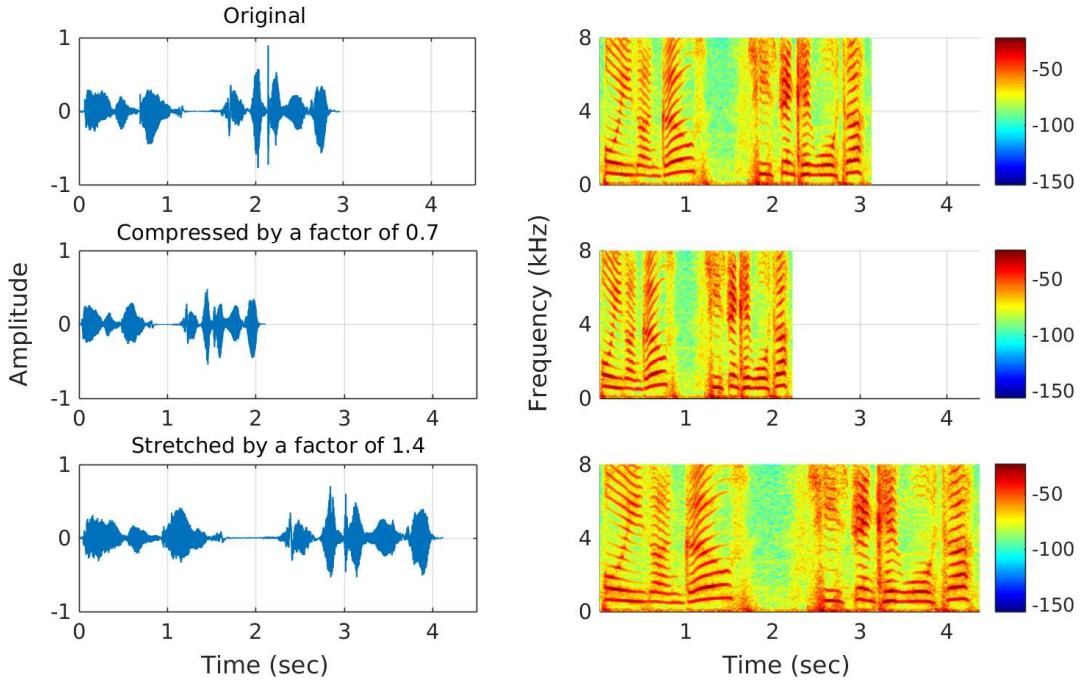


Fig. 2: The time-domain waveforms for a segment of speech and the corresponding compressed and stretched versions. The spectrograms for each of the cases are also shown. It is noted that, the shape of the signal as well as formant transitions remain preserved even after time-scale modification.

increasing/decreasing the speaking-rate. Next, the speaking-rate of the children’s speech used for the analysis presented in Figure 1 was increased by modifying the duration of each of the utterances using a factor of 0.7. The histogram depicting the variation of speaking-rate after modification are shown in Figure 3 (bottom pane). For proper contrast, the histogram obtained prior to TSM is also shown Figure 3 (top pane). By comparing the two histograms, an increase in mean speaking-rate is noticeable. It can be concluded from these analyses that, by optimally compressing the signal duration, the acoustic mismatch resulting from the differences in speaking-rate can be reduced to a large extent. In the next section, the simulation studies performed to validate this claim are presented in detail.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, we first describe the experimental setup employed in this work. This is followed by the experimental evaluations validating the effectiveness of speaking-rate adaptation.

A. Experimental setup

Speech Corpora: Adults’ speech data used in this work was obtained from WSJCAM0 [16] British English speech corpus for continuous speech recognition. A train set (Adult-Train) was derived from WSJCAM0 for learning the statistical model parameters. Adult-Train set consisted of 15.5 hours of speech data from 92 adult male and female speakers. Further, the train set comprised of 132,778 words and the total number of utterances was 7852. A test set (Adult-Test) was also derived in order to measure the matched case recognition

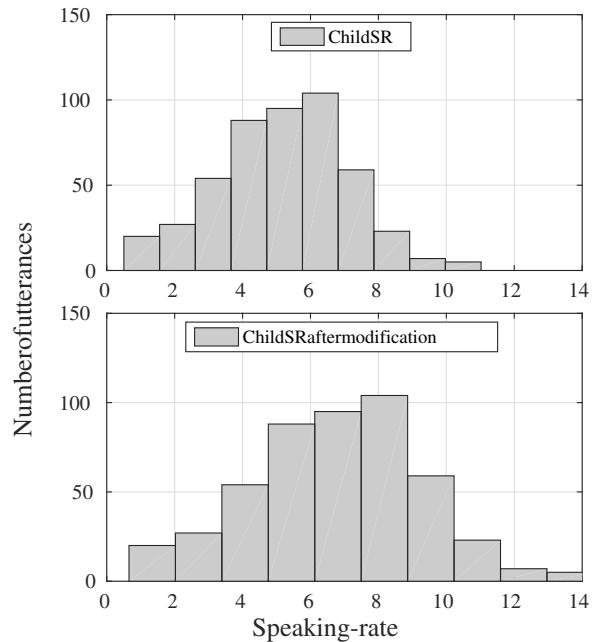


Fig. 3: Histogram depicting variation in speaking-rate for children before and after time-scale modification. Speaking-rate is quantified in terms of number of phonemes per second

performance. The Adult-Test set consisted of 0.6 hours of speech data from 20 speakers with a total of 5,608 words. In order to simulate the mismatched ASR task discussed earlier, children’s speech data was obtained from PF-STAR corpus [17]. Like WSJCAM0, PF-STAR corpus is also a British English speech database. A test set (Child-Test) was

derived from PF-STAR corpus which consisted of speech data from 60 children aged between 3 to 14 years. The total duration of speech data was 1.1 hours. There were a total of 5067 words in Child-Test data set. The experimental studies reported in this work were performed on a wide-band speech data sampled at a rate of 16 kHz.

Front-end speech parameterization: The speech data was first pre-emphasized using a high-pass filter. The pre-emphasis factor was selected as 0.97. Next, frame-blocking was done using overlapping Hamming windows of length 20 ms with an overlap of 50%. In other words, the frame-overlap was chosen to be 10 ms. In order to extract the 13-dimensional base Mel-frequency cepstral coefficients (MFCC), a 40-channel Mel-filter bank was employed. The 13-dimensional base MFCC features were then spliced in time taking a context size of 9 frames. Time-splicing resulted in 117-dimensional vectors which were then reduced to 40-dimensional features vectors using linear discriminant analysis (LDA) and maximum-likelihood linear transformation (MLLT). Cepstral mean and variance normalization (CMVN) as well as feature-space maximum-likelihood linear regression (fMLLR) were performed next to enhance the robustness towards speaker-dependent variations. The required fMLLR transformations for the training and test data were generated through speaker adaptive training.

Specifications of the ASR system: The ASR systems were developed on the 15.5 hours adults' speech data from the WSJCAM0 speech corpus using the Kaldi toolkit [22]. Context-dependent hidden Markov models (HMM) were used for modeling the cross-word triphones. Decision tree-based state tying was performed with the maximum number of tied-states (senones) being fixed at 2000. Acoustic modeling based on deep neural network and the long short-term-memory recurrent neural network (RNN) was explored as stated earlier. Prior to learning parameters of the DNN-HMM-based ASR system, the fMLLR-normalized feature vectors were time-spliced once again considering a context size of 9 frames. The number of hidden layers in the DNN was chosen as 8 with each layer consisting of 1024 hidden nodes. The nonlinearity in the hidden layers was modeled using the *tanh* function. The initial learning rate for training the DNN-HMM parameters was set at 0.005 which was reduced to 0.0005 in 15 epochs. The minibatch size for neural net training was selected as 512. The LSTM-based acoustic models were trained with 4 hidden layers each having 1024 nodes. The dimension of the LSTM cell was chosen as 1024. The number of epochs used for LSTM training was set to 5 while the initial and final learning rates were selected to be 0.005 and 0.0005, respectively.

While evaluating the matched case performance or decoding the Adult-Test set, the MIT-Lincoln 5k Wall Street Journal bi-gram language model (LM) was used. The perplexity of this LM for the Adult-Test set is 95.3 while there are no out-of-vocabulary (OOV) words. Further, a lexicon consisting of 5,850 words including pronunciation variants was used. While decoding the Child-Test set, a 1.5k domain-specific bigram LM was used. This bigram LM was trained on the transcripts of speech data in PF-STAR after excluding those

TABLE I: Baseline WERs for Child-Test and Adult-Test sets on adult data trained DNN- and LSTM-based ASR systems.

Acoustic modeling technique	WER (in %)		Relative difference (%)
	Child-Test	Adult-Test	
DNN	19.27	5.87	70
LSTM	16.33	5.10	69

corresponding to Child-Test set. The said domain-specific LM has an OOV rate of 1.20% and perplexity of 95.8 for the Child-Test set. The lexicon used while decoding the Child-Test set consisted of 1,969 words including pronunciation variations.

B. Baseline recognition performances

The word error rate (WER) metric was employed to measure the recognition performance. The baseline WERs for Child-Test and Adult-Test datasets with respect to the adult data trained DNN- and LSTM-based ASR systems are given in Table I. The ill-effects of aforementioned factors of acoustic mismatch can be easily understood by noting the large difference in WERs for the two sets. It is to note that, the presented WERs were obtained after applying CMVN and fMLLR in order to reduce the speaker-dependent acoustic mismatch. In addition to that, domain-specific LMs were used while decoding the corresponding test sets. Yet, the recognition performance for children's speech is much poorer compared to adults' case. The relative difference in WERs for the two test sets highlight this point.

C. Effect of speaking-rate adaptation

In order to change the speaking-rate for children's speech, the TSM factor was varied from 0.65 to 1.35 in steps of 0.05. Modification factor values less than unity imply an increase in speaking-rate. On the other hand, in order to decrease the speaking-rate, values greater than one are used. The correspondingly modified test data was then decoded to improve the recognition rates. The WER profiles demonstrating the effect increasing and decreasing the speaking-rate are shown in Figure 4. Since the speaking-rate is lower in the case children's speech, the WER is observed to decrease when the scaling factor is less than unity. Similar trends are noted for both DNN- as well as LSTM-based ASR systems. The best case WERs along with the percentage relative improvement over the baseline obtained through SRA are given in Table II. From these results it is evident that proposed approach for SRA is extremely effective.

IV. CONCLUSION

The role of speaking-rate adaptation in the context of automatic speech recognition has been studied in this work. In this regard, a recently proposed time-scale modification technique based on fuzzy classification of spectral bins is explored. This fuzzy-classification-based TSM approach is reported to be better than the existing similar techniques. In order to simulate

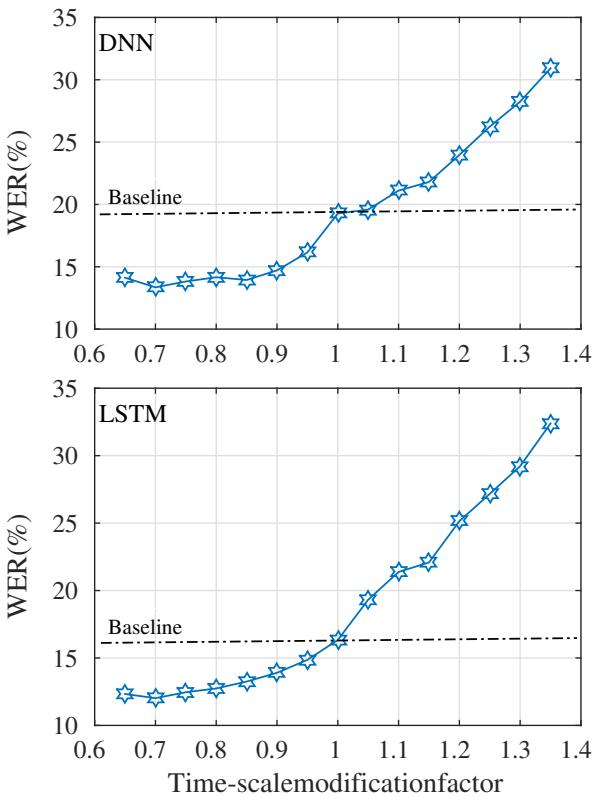


Fig. 4: WERs illustrating the effect of increasing/decreasing the speaking-rate on the recognition of children's speech using DNN- and LSTM-based ASR system trained on adults' speech.

TABLE II: The best case WERs for children's speech test set obtained through SRA.

Acoustic model	WER (in %)		Relative improvement (%)
	Baseline	SRA	
DNN	19.27	13.35	30.7
LSTM	16.33	12.02	26.4

an ASR task where large differences in speaking-rate exists, children's speech is transcribed using acoustic models trained on speech data from adult speakers. Significant reductions in word-error rates are obtained by suitably changing the speaking-rate through fuzzy-classification-based TSM.

REFERENCES

- [1] J. Driedger and M. Müller, "A review of time-scale modification of music signals," *Applied Sciences*, vol. 6, no. 2, p. 57, 2016.
- [2] R. Kent and L. Forner, "Speech Segment Durations in Sentence Recitations by Children and Adults," *Journal of Phonetics*, vol. 8, pp. 157–168, 1980.
- [3] S. Lee, A. Potamianos, and S. S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. INTERSPEECH*, vol. 1, September 1997, pp. 473–476.
- [4] ———, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, March 1999.
- [5] S. M. Chu and D. Povey, "Speaking rate adaptation using continuous frame rate normalization," in *Proc. ICASSP*, March 2010, pp. 4306–4309.
- [6] M. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech," in *Proc. Speech and Language Technologies in Education (SLaTE)*, September 2007.
- [7] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. Workshop on Child, Computer and Interaction*, 2009, pp. 7:1–7:8.
- [8] E.-P. Damaskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, no. 12, p. 1293, 2017.
- [9] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic normalization of childrens speech," in *Proc. INTERSPEECH*, September 2003, pp. 1313–1316.
- [10] S. Ghai, "Addressing Pitch Mismatch for Children's Automatic Speech Recognition," Ph.D. dissertation, Department of EEE, Indian Institute of Technology Guwahati, India, October 2011.
- [11] G. E. Hinton, L. Deng, D. Yu, G. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.
- [12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [13] J. L. Miller, "Effects of speaking rate on segmental distinctions," *Perspectives on the study of speech*, pp. 39–71, 1981.
- [14] Q. Summerfield, "Articulatory rate and perceptual constancy in phonetic perception," *Journal of Experimental Psychology: Human Performance and Perception*, vol. 7, pp. 208–215, 1981.
- [15] J. L. Miller and L. E. Volaitis, "Effect of speaking rate on the perceptual structure of a phonetic category," *Perception & Psychophysics*, vol. 46, no. 6, pp. 505–512, November 1989.
- [16] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition," in *Proc. ICASSP*, vol. 1, May 1995, pp. 81–84.
- [17] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, and M. Wong, "The PF_STAR children's speech corpus," in *Proc. INTERSPEECH*, 2005, pp. 2761–2764.
- [18] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proc. INTERSPEECH*, 2005, pp. 1137–1140.
- [19] T. S. Verma and T. H. Meng, "An analysis/synthesis tool for transient signals that allows a flexible sines+ transients+ noise model for audio," in *Proc. ICASSP*, vol. 6, 1998, pp. 3573–3576.
- [20] L. A. Zadeh, "Making computers think like people," *IEEE spectrum*, vol. 21, no. 8, pp. 26–32, 1984.
- [21] J. Driedger, M. Müller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.
- [22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, December 2011.

Instantaneous Frequency Features for Noise Robust Speech Recognition

Shekhar Nayak

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India
ee13p1008@iith.ac.in*

Shashank Dhar B.

*School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
shashankdhar@gatech.edu*

Saurabhchand Bhati, Koilakuntla Bramhendra and K. Sri Rama Murty

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India
ee12b1044@iith.ac.in, ee17mtech01002@iith.ac.in, ksrm@iith.ac.in*

Abstract—Analytic phase of the speech signal plays an important role in human speech perception, specially in the presence of noise. Generally, phase information is ignored in most of the recent speech recognition systems. In this paper, we illustrate the importance of analytic phase of the speech signal for noise robust automatic speech recognition. To avoid phase wrapping problem involved in the computation of analytic phase, features are extracted from instantaneous frequency (IF) which is time derivative of analytic phase. Deep neural network (DNN) based acoustic models are trained on clean speech using features extracted from the IF of speech signals. Robustness of IF features in combination with mel-frequency cepstral coefficients (MFCCs) was evaluated in varied noisy conditions. System combination using minimum Bayes risk decoding of IF features with MFCCs delivered absolute improvements of upto 13% over MFCC features alone for DNN based systems under noisy conditions. The impact of the system combination of magnitude and phase based features on different phonetic classes was studied under noisy conditions and was found to model both voiced and unvoiced phonetic classes efficiently.

I. INTRODUCTION

In the recent years, there has been significant advancement in the performance of automatic speech recognition (ASR) systems. Notwithstanding with this progress, there is a wide gap between the performance of humans and ASR systems in real-life environments. Studies suggest that humans can understand speech in unseen noisy environments without any training [1]. Whereas, ASR systems tend to perform poorly in noisy environments. Specially, with lower signal-to-noise ratio (SNR) values, the systems trained on traditional magnitude based features on clean data are rendered incapable of estimating noisy data distributions and therefore require training with noisy speech [2].

An alternate approach to training with noisy data is to explore noise robust features for speech recognition which can perform well in clean as well as unknown noisy conditions. Several features have been proposed for noise robust speech

recognition. Dimitriadis *et al.* used AM-FM features in combination with MFCCs [3] to train Gaussian mixture models (GMM) - hidden Markov model (HMM) based recognition systems on clean speech and conducted tests under different noisy conditions. These hybrid features provided good performance in clean as well as mismatched conditions with different noises at various SNRs. High resolution temporal envelopes and sub-band envelopes derived from frequency domain linear prediction of speech have been used for noise robust phoneme recognition and also for recognizing broad phonetic classes [4], [5]. Features from Gammatone wavelet filterbanks showed improvements in recognition accuracy at low SNRs compared to MFCC features [6]. Envelope of Gammatone filterbank output along with IF were used in clean and noisy conditions for ASR [7]. Fourier-Bessel cepstral coefficients derived from zeroth order Bessel functions have been used for recognition in noisy conditions [8]. Tonal features along with MFCC features provided better recognition in different noisy conditions with different acoustic models [9].

The importance of analytic phase of speech was established through perceptual studies [10] which showed that analytic phase of speech signal improves human speech recognition in noise. It has also been shown that uncertainty in phase results in higher word error rates or lower intelligibility for human listeners [11]. Use of dynamic frequency modulated (FM) systems improved speech recognition in noise for cochlear implant recipients [12]. These works strongly demonstrate the indispensable nature of phase information in human perception.

A prominent reason for less exploration of analytic phase in speech recognition is the phase wrapping problem. It is difficult to discriminate between phase values differing by integer multiples of 2π . Therefore, it is computationally expensive to extract features from the phase component as compared to the magnitude component. Since, there has been no dearth of computational resources in recent times, this has lead to

resurgence of phase based methods in speech community [13].

There have been several previous approaches to incorporate different IF based features along with magnitude features to enhance ASR performance in noisy conditions. The output of an array of band-pass filters was decomposed into analytic and anti-analytic components to derive average IF and average log envelope features for clean training to recognize noisy speech [14]. Mean instantaneous amplitude, mean IF and frequency modulation percentages were used in addition to MFCC features for training GMM-HMM based recognition systems on clean speech and testing under various noisy conditions [3]. Temporal AM and FM features were used for speech recognition in noisy environments [15]. Sub-band IF features with wavelet sub-band features were used for phoneme recognition tasks on TIMIT database under noisy conditions [16]. Group delay based features and features from modified group delay function were used for several recognition tasks independently and in conjunction with MFCC features in clean and noisy conditions [17], [18]. Group delay features derived from all pole models were used for speaker recognition [19]. Robust syllable based segmentation and recognition on TIMIT data has been shown in [20]. Different IF feature extraction methods were investigated for phoneme recognition on clean speech [21].

In this paper, we discuss the performance of features extracted from smoothed IF, derived from the analytic phase of speech signal for noise robust speech recognition. The rest of the paper is organized as follows: Section II discusses feature extraction from IF of speech signals, Section III describes recognition experiments using IF features on clean and noisy data, Section IV consists of experimental results and Section V concludes the paper.

II. INSTANTANEOUS FREQUENCY FEATURE EXTRACTION

The significance of phase of the speech signal becomes higher in noisy conditions as the speech recognition performance becomes worse in lower SNRs [11]. Therefore, in this work phase based instantaneous frequency features are explored to complement magnitude based features for improving recognition performance in noisy conditions. Feature extraction from instantaneous frequency of speech signals is discussed below.

For a real signal $s(t)$, the corresponding complex analytic signal $s_a(t)$ can be expressed as

$$s_a(t) = s(t) + j s_h(t), \quad (1)$$

where the Hilbert transform of $s(t)$ is denoted by $s_h(t)$ [22]. The polar form of the analytic signal in (1) is given by

$$s_a(t) = a(t)e^{j\phi(t)} \quad (2)$$

where $a(t)$ denotes the amplitude envelope and $\phi(t)$ denotes the analytic phase of the signal. Instantaneous frequency (IF) is defined as the time derivative of the unwrapped analytic phase $f_i(t)$ [22], and is given by

$$\Psi(t) = \frac{1}{2\pi} \frac{d\phi(t)}{dt} \quad (3)$$

The IF of a narrow-band discrete signal $s[n]$ can be computed by first-order differencing the unwrapped phase. For the computation of unwrapped phase, there are numerous phase unwrapping methods but most of them are adhoc and inaccurate [23].

To compute the IF, the logarithm of (2) is differentiated with respect to t and the imaginary parts are equated as below

$$f_i(t) = \frac{1}{2\pi} \phi'(t) = \frac{1}{2\pi} \Im \left\{ \frac{s'_a(t)}{s_a(t)} \right\} \quad (4)$$

where $s_a(t)$ and $s'_a(t)$ denote the analytic signal and its derivative, respectively. $\Im \{\cdot\}$ denotes the imaginary component. $s'_a(t)$ is computed by the differentiation property of Fourier transform. In discrete domain, IF can be implemented as [24]

$$f_i[n] = \frac{1}{N} \frac{\Im \{s'_a[n]s_a[n]\}}{|s_a[n]|^2} \quad (5)$$

where N denotes the length of the analytic signal, $|s_a[n]|^2$ denotes its amplitude envelope and $s'_a[n]$ denotes its first derivative. $s'_a[n]$ can be computed by applying the differentiation property of discrete Fourier transform (DFT) as

$$s'_a[n] = j\mathcal{F}^{-1} \{kS_a[k]\} \quad (6)$$

where $S_a[k]$ denotes DFT of $s_a[n]$ and \mathcal{F}^{-1} denotes inverse DFT.

IF is generally defined for narrow-band signals whereas speech signal $s[n]$ is a wide-band signal. Therefore, we pass the speech signal through a filter-bank of 40 linear, Gaussian shaped narrow-band filters, with 400 Hz bandwidth, centered at $(p-1) \times 200$ Hz, for $p = 1, 2, \dots, 40$. Through this operation, 40 narrow-band components are obtained from the speech signal, $s_i[n]$, $i = 1, 2, \dots, 40$. IF can be computed from each $s_i[n]$ as $f_i[n]$, $i = 1, 2, \dots, 40$ using (5) after smoothing the numerator and denominator separately.

Zero-mean IF is obtained by subtracting each filter's centre frequency from the IF corresponding to that filter. Then, IF is averaged over every 25 ms frame with a frame shift of 10 ms. Hence, a 40-dimensional feature vector is obtained from average IF values corresponding to 40 filters. On this feature vector, discrete cosine transform (DCT) is applied and first 13 coefficients are retained. These features are referred as instantaneous frequency cosine coefficients (IFCCs) [25]. A 39-dimensional feature vector is obtained by appending their first and second order time derivatives. These features have been successfully applied to improve speaker verification performance along with MFCC features [26].

Fig. 1(a) shows a clean speech signal from TIMIT database, Fig. 1(b) shows the Spectrogram for this signal and Fig. 1(c) shows Pyknogram [27] of Smoothed IFCC features extracted using 64 such filters. Fig. 1(d), Fig. 1(e) and Fig. 1(f) show corresponding figures for a noisy version of the same utterance with 10 dB white noise. The Pyknogram clearly exhibits the variations in formants and slight degradation is visible in case of noisy version.

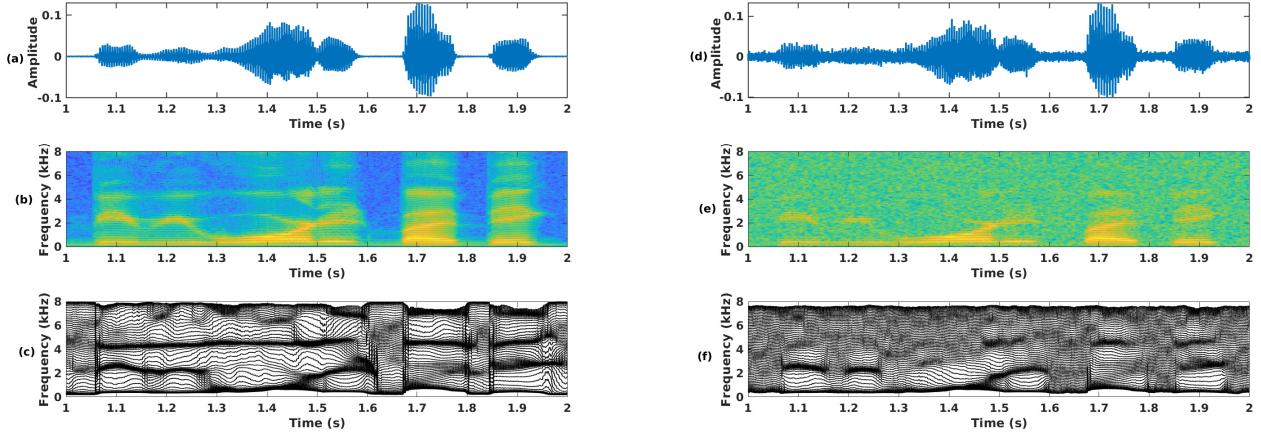


Fig. 1. (a) Speech signal, (b) Spectrogram, and (c) Pyknogram of smoothed IFCC features for clean utterance from TIMIT. (d) Speech signal, (e) Spectrogram, and (f) Pyknogram of smoothed IFCC features for the same utterance with 10 dB white noise.

III. NOISY SPEECH RECOGNITION USING IF FEATURES

A. Database

The speech recognition experiments in this paper are conducted on TIMIT database [28]. The data is divided into training, development and evaluation sets with 462, 50 and 24 speakers respectively. The hybrid DNN-HMM based recognition systems and GMM-HMM baseline systems are trained on core training set consisting of 3696 sentences. The development set and the test set consists of 400 sentences and 192 sentences, respectively. We created noisy version of TIMIT dev and test sets by adding white, babble and car noise at 10 dB, 15 dB and 20 dB SNR levels. Noise samples are taken from noise files provided with [29]. A bigram language model is used during recognition.

B. GMM-HMM clean training

MFCC features [30] for the GMM-HMM baseline system are obtained as follows. Short-time Fourier transform (STFT) is performed on overlapping hamming windowed frames of 25 ms frame size and 10 ms frame shift. MFCC features are obtained by applying DCT to log Mel-scale filter-bank outputs and choosing first 13 coefficients. IF features are extracted using filter-bank consisting of 40 filters with 3-dB bandwidth of 400 Hz. Again, DCT is applied and first 13 coefficients are retained as IFCCs.

The features for training GMM-HMM baseline system are obtained as follows. Splicing is done for 7 frames (3 left-context, 1 current and 3 right-context) of 39-dimensional MFCCs (with delta and acceleration coefficients) and the resultant is projected to 40 dimensional feature vector using linear discriminant analysis (LDA) and semi-tied covariance (STC) transform is applied [31]. Further, speaker adaptive training (SAT) is done on these features using a single feature-space maximum likelihood linear regression (FMLLR) transform estimated per speaker [32]. The features thus obtained are used

to train the baseline GMM-HMM systems. For the trained system, the number of tied triphone states (senones) are 1951.

C. DNN-HMM clean training

The DNNs in the experiments consists of 6 hidden layers with 1024 neurons (sigmoid activation) in each hidden layer. The total number of output units are 1951 for each DNN. The input to the DNN is 11 frames (5 left-context, 1 current and 5 right-context) of 40 dimensional features concatenated together. Restricted Boltzmann Machine (RBM) pre-training is done in a greedy layerwise fashion [33]. DNNs initialized from the resultant deep belief network are trained by mini-batch stochastic gradient descent [34]. Targets for DNN are obtained by forced alignment using the above GMM-HMM system.

Separate DNN-HMM systems with above configuration are trained on MFCC and IF features extracted from clean training set. The targets are common for both the DNNs using MFCC and IF features. All the features are speaker adapted using FMLLR before being used as inputs for DNNs. The systems are trained using the Kaldi speech recognition toolkit [35].

The DNN-HMM based recognizers trained on MFCC features exploit the phonetic information from only the magnitude spectrum of the speech where as the recognizers trained on IFCC features source the phonetic information from the phase spectrum only. In order to benefit from the complimentary phonetic information from the phase spectrum along with the magnitude spectrum, we propose to combine the two different systems at the scoring level.

D. System combination of MFCC and IFCC features

Lattices are combined for DNN-HMM based systems trained on MFCC and IF features using Minimum Bayes Risk (MBR) decoding [36]. MBR decoding is an approach to combine systems using multiple lattices by minimizing the expected phone error rate (PER). Each lattice is provided

TABLE I
COMPARISON OF PHONE ERROR RATES (%) FOR DIFFERENT MFCC BASED MONOPHONE GMM-HMM SYSTEMS (BASELINE) AND THEIR CORRESPONDING COMBINED SYSTEMS WITH DIFFERENT IF FEATURES AT 10 dB SNR.

System	TIMIT	TIMIT + White	TIMIT + Babble	TIMIT + Car
MFCC ¹ [3]	41.6	82.28	72.29	47.25
MFCC (Proposed)	32.6	70.4	56.9	37.5
MFCC + IF ¹ [3]	40.59	73.95	61.4	43.5
MFCC+IFCC (Proposed)	32.2	66.0	57.6	43.4

TABLE II
PHONE ERROR RATES (%) IN CLEAN AND NOISY CONDITIONS FOR DNN-HMM SYSTEMS TRAINED ON CLEAN SPEECH.

System	Clean	White			Babble			Car Noise		
		10	15	20	10	15	20	10	15	20
MFCC	18.8	66.9	56.1	44.7	48.3	37.8	29.2	27.7	25.8	29
IFCC	22.1	69.1	53.8	43.1	51.0	41.1	34.1	46.8	42.5	36.6
MFCC+IFCC	17.3	53.9	44.4	35.5	40.6	32.5	26.5	29.7	27.3	24.5

TABLE III
PHONE ERROR RATES (%) FOR PHONETIC CLASSES - VOWELS, FRICATIVES AND PLOSIVES FOR WHITE NOISE.

System	Vowels			Fricatives			Plosives		
	10	15	20	10	15	20	10	15	20
MFCC	53.5	42.0	34.8	89.3	72.6	53.1	98.6	91.4	73.6
IFCC	60.4	46.5	37.4	98.9	73.8	54.0	94.3	83.2	67.3
MFCC+IFCC	52.9	41.9	34.5	78.3	65.8	46.6	96.8	87.5	72.5

weights in the system combination. Optimal phone sequence is decoded based on the Levenshtein distance between two phone sequences from different lattices. Equal weights were considered for MFCC and IFCC features based lattices for all the experiments.

IV. EXPERIMENTAL RESULTS

The phoneme recognition performance is evaluated in terms of PER on clean TIMIT core test set and noisy TIMIT test sets with white, babble and car noises at different SNRs. Table I shows the comparison of PERs for MFCC based monophone GMM-HMM systems from [3] and current work for the above noises and clean conditions at 10 dB SNR. Also, PERs are evaluated for corresponding MFCC based systems combined with IF-Mean features from [3] and IFCC features from current work. The difference in the MFCC baselines can be attributed to the difference in monophone training strategies. The system combination of proposed features with MFCC provided absolute improvement of 8.39%, 7.95%, 3.8% and 0.1% for clean, white noise, babble noise and car noise conditions respectively over the MFCC+IF-Mean system in [3]. Therefore, the IFCC features provide significantly better combination with MFCCs for recognition on clean speech and with white noise compared to mean IF features.

Table II shows the results for DNN-HMM systems trained on clean speech and tested in different noise conditions at different SNR levels for MFCC, IFCC and their system combinations. The system trained with MFCC features from clean speech recorded a PER of 18.8% on TIMIT core test

set which is better than 20.0% PER of the best system reported in [34] based on convolutional DNNs trained with Mel filter-bank features. The system combination of MFCC and IF features provided absolute improvement of 13%, 11.7% and -2% over MFCC alone for white noise, babble noise and car noise respectively at 10 dB SNR. The improvements reduced to 9.2%, 2.7% and 4.5% at 20 dB. This shows the robustness of IFCC features in more noisy conditions. Also, IFCC features provide the highest performance improvement for speech with white noise. As the SNR increases, recognition accuracy improves as it is closer to matched training conditions and hence there is reduction in improvement at higher SNRs. There is slight degradation in performance of the combined system in car noise conditions at lower SNRs as MFCCs are able to model significantly better than IFCCs in this case. But, the combination performs better than MFCCs as the SNR increases to 20 dB. The performance of IFCC features and combined systems for different phonetic classes in white noise conditions is given in Table III. Three broad phonetic classes - vowels, fricatives and plosives are considered for this evaluation. Semi-vowels have also been considered into the vowel category. There is significant improvement of 11%, 6.8% and 6.5% in case of fricatives at 10 dB, 15 dB and 20 dB SNR levels. Slight improvement is observed in case of vowels and plosives as well. This shows that IFCC features along with MFCCs can recognize different phonetic classes both voiced and unvoiced in a better way than only MFCC features in both clean and noisy conditions.

V. CONCLUSION

This paper investigates instantaneous frequency features for noise robust speech recognition. Recognition experiments

¹The phoneme accuracies given in the paper are converted to PERs for comparison.

were conducted for TIMIT phone recognition task under clean and various noisy conditions. The magnitude and IF features based GMM-HMM and DNN-HMM systems were trained on clean speech. MBR decoding was used to combine MFCC and IFCC based systems. System combination of both features delivered absolute improvements of upto 13% over MFCC features alone for DNN-HMM systems under noisy conditions. IFCC features in combination with MFCC features provided significant improvement for all phonetic classes in clean and white noise conditions. The improvement was significantly higher for unvoiced phones. IFCC features are more robust in higher noise levels. This work demonstrates the significance of combining evidences from magnitude and phase for noise robust speech recognition. IF based features are effective in conjunction with magnitude based features for different types of noises, different levels of noise and also in different broad phonetic classes.

REFERENCES

- [1] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.
- [2] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87*, vol. 12. IEEE, 1987, pp. 705–708.
- [3] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust am-fm features for speech recognition," *IEEE signal processing letters*, vol. 12, no. 9, pp. 621–624, 2005.
- [4] S. Ganapathy and H. Hermansky, "Robust phoneme recognition using high resolution temporal envelopes," in *Proc. of INTERSPEECH*, 2012.
- [5] S. Ganapathy, S. Thomas, and H. Hermansky, "Modulation frequency features for phoneme recognition in noisy speech," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. EL8–EL12, 2009.
- [6] A. Adiga, M. Magimai, and C. S. Seelamantula, "Gammatone wavelet cepstral coefficients for robust speech recognition," in *TENCON 2013-2013 IEEE Region 10 Conference (31194)*. IEEE, 2013, pp. 1–4.
- [7] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech communication*, vol. 53, no. 5, pp. 707–715, 2011.
- [8] C. Prakash and S. V. Gangashetty, "Fourier-bessel cepstral coefficients for robust speech recognition," in *Signal Processing and Communications (SPCOM), 2012 International Conference on*. IEEE, 2012, pp. 1–5.
- [9] A. Dey, B. D. Sarma, W. Lalhminglui, L. Ngente, P. Gogoi, P. Sarmah, S. Prasanna, R. Sinha, and S. Nirmala, "Robust mizo continuous speech recognition."
- [10] F.-G. Zeng, K. Nie, G. S. Stickney, Y.-Y. Kong, M. Vongphoe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 7, pp. 2293–2298, 2005.
- [11] G. Shi, M. M. Shafechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 5, pp. 1867–1874, 2006.
- [12] J. Wolfe, E. C. Schafer, B. Heldner, H. Mülder, E. Ward, and B. Vincent, "Evaluation of speech recognition in noise with cochlear implants and dynamic fm," *Journal of the American Academy of Audiology*, vol. 20, no. 7, pp. 409–421, 2009.
- [13] P. Mowlaei, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.
- [14] Y. Wang, J. Hansen, G. K. Allu, and R. Kumaresan, "Average instantaneous frequency (aif) and average log-envelopes (ale) for asr with the aurora 2 database," in *INTERSPEECH*, 2003.
- [15] Y. Kubo, S. Okawa, A. Kurematsu, and K. Shirai, "Temporal am-fm combination for robust speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 716–725, 2011.
- [16] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Feature extraction technique using erb like wavelet sub-band periodic and aperiodic decomposition for timit phoneme recognition," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 389–399, 2014.
- [17] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 190–202, 2007.
- [18] R. Ramya, R. M. Hegde, and H. A. Murthy, "Significance of group delay based acoustic features in the linguistic search space for robust speech recognition," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.
- [19] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alku, "Using group delay functions from all-pole models for speaker recognition." in *INTERSPEECH*. Citeseer, 2013, pp. 2489–2493.
- [20] R. Janakiraman, J. C. Kumar, and H. A. Murthy, "Robust syllable segmentation and its application to syllable-centric continuous speech recognition," in *Communications (NCC), 2010 National Conference on*. IEEE, 2010, pp. 1–5.
- [21] S. Nayak, S. Bhati, and K. S. R. Murty, "An investigation into instantaneous frequency estimation methods for improved speech recognition features," in *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on*. IEEE, 2017, pp. 363–367.
- [22] C. Leon, "Time-frequency analysis: theory and applications," USA: Prentice Hall, 1995.
- [23] E. Loweimi, S. M. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International conference on*, 2013, pp. 7155–7159.
- [24] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.
- [25] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analytic phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.
- [26] K. Vijayan, V. Kumar, and K. S. R. Murty, "Feature extraction from analytic phase of speech signals for speaker verification," in *INTERSPEECH*, 2014, pp. 1658–1662.
- [27] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.
- [28] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *Speech Input/Output Assessment and Speech Databases*, 1989.
- [29] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.
- [30] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [31] M. J. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE transactions on speech and audio processing*, vol. 7, no. 3, pp. 272–281, 1999.
- [32] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance gaussians," in *INTERSPEECH*, 2006.
- [33] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.
- [34] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [35] D. Povey, A. Ghoshal, G. Boulian, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.
- [36] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech & Language*, vol. 25, no. 4, pp. 802–828, 2011.

Slit Loaded Circular Ultra Wideband Antenna for Band Notch Characteristics

Ameya A Kadam
Research Scholar,
Dept of EXTC,
SVKM's DJSCOE
Mumbai, India
ameyakadam@rediffmail.com

Amit A Deshmukh
Professor & Head,
Dept of EXTC,
SVKM's DJSCOE
Mumbai, India
amitdeshmukh76@rediffmail.com

S B Deshmukh
Research Scholar,
Dept of EXTC,
SVKM's DJSCOE
Mumbai, India
sanjay.deshmukh@djsce.ac.in

Akshay Doshi
PG Student,
Dept of EXTC,
SVKM's DJSCOE
Mumbai, India
doshiakshay4192@gmail.com

Kamla Prasan Ray
Professor & Head,
Dept of ETRX,
DIAT
Pune, India
kpray@rediffmail.com

Abstract—This paper proposed a planar, low cost, simple, and compact printed microstrip-fed circular monopole ultra-wideband antennas with band-notched characteristics. By introducing pair of open slits and varying angular separation between slits on the circular patch the band notched characteristics can be obtained. The proposed antennas are successfully simulated, designed, fabricated on FR-4 substrate. The measured results show that the proposed antenna with dimensions of 65mm × 65 mm × 1.6mm has a bandwidth over the frequency band 2.00–10.6 GHz with VSWR ≤ 2, except 2.5–3.95GHz with circular antenna having pair of slits. The presented antennas shows nearly omnidirectional radiation pattern, stable gain, small group delay variation at working frequencies. Satisfactory results have been obtained in frequency and time-domain analysis of the proposed structure.

Keywords—circular patch, tunable notch characteristics, ultrawideband printed antenna, slit loaded antenna

I. INTRODUCTION

Microstrip antenna (MSA) can be easily integrated with microwave integrated circuit (MIC) and hence it is generally used in high frequency circuit. Hence microstrip variant of planar monopole antennas are commonly used to cover ultrawide band (UWB) which occupies a frequency range between 3.1 and 10.6 GHz as approved by the Federal Communications Commission (FCC) in 2002. The regularly used UWB antennas with patch shapes are circular, rectangular, triangular, hexagonal and their amended variations [1–6]. The UWB antennas are designed by cutting various shapes of slots in the radiating patch or in the ground plane. [6–8]. Other narrowband services like, WiMAX 802.16 (3.3–3.8 GHz), IEEE 802.11a exist over the designated UWB spectrum (3.1–10.6 GHz). For some applications, UWB antenna along with additional filters are used to reject these bands which increases the complication of UWB system and cost. Hence, to reduce the interferences between narrowband systems and UWB, it is more appropriate to realize UWB antennas with notched frequency bands. Numerous designs of antenna have been reported for notch frequency response in ultra wideband. [7–11]. As per reported literature, introduction of stubs, slots in patch or ground plane results into notch band characteristics. However, detailed explanation for highlighting the effects of modifications in patch geometry that yields tunable notch response in terms of patch resonant modes is not provided.

In this paper, an innovative design of antenna having circular shaped with pair of slits is proposed. First, the UWB antenna with circular shape is designed. An input impedance (Z_{in}) response for magnitude of $S_{11} < -10\text{dB}$, for the frequencies from <1.8 GHz to >10 GHz is obtained. The

circular UWB antenna yields peak gain of around 2 dBi with nearly omnidirectional radiation pattern. Further using optimized circular shaped UWB antenna along with pair of slits, a notch characteristic over frequency range of 1.9–10.3 GHz is realized. The circular shape antenna is loaded with pair of slits and an angular separation between slits is varied to obtain tunable notch characteristics. These modifications in circular shape patch results into alteration of the resonance frequencies and input impedance for the resonant modes across 2–10 GHz frequency range which yields notch characteristics. The placement of the slits perturbs the resonant frequency and impedance of TM_{21} mode significantly. The tuning of notch band is provided either by varying the length of the slits (L_s) or by changing the angular separation (θ) of the slits. The return loss (S_{11}) of more than -5 to -6 dB is obtained within the notch band, ensuring more than 45% of reflected power. Within the pass band, circular shape antenna with pair of slits shows nearly omnidirectional radiation pattern with gain of around 1–1.5 dBi. The proposed antenna configurations were optimized using IE3D simulations on low cost FR4 substrate ($\epsilon_r = 4.4$, $h = 1.6$ mm, $\tan \delta = 0.02$), followed by the measurements. The input impedance response upto 8 GHz is measured using vector network analyzer (ZVH – 8) whereas the broadside co-polar gain and radiation pattern are measured using spectrum analyzer (FSC 6) and RF source (SMB 100A). The proposed UWB antennas can be useful for applications in Bluetooth, Wi-MAX and Wi-Fi applications in 2–10 GHz frequency range. The use of tunable band notch using modified patch can help into minimization of an interference with parallel applications.

II. ULTRA WIDEBAND ANTENNA WITH CIRCULAR SHAPE

The circular shape antenna is designed at lower cut-off frequency of around 1.8GHz on FR-4 substrate with height $h = 1.59$ mm, dielectric permittivity $\epsilon_r = 4.3$, and loss tangent $\tan \delta = 0.02$. The geometry of the patch is shown in Figure 1(a). The radius of the circular patch is $r = 14$ mm. A microstrip line feed having width of 3 mm with partial ground plane with dimensions $L_g = 65$ mm and $W_g = 20$ mm, has been used to feed the antenna. The simulated resonance curve plots for circular patch is shown in Figure 1(b). For the frequencies from 2–10 GHz the simulated impedance response for return loss less than -10 dB is shown in Figure 1b. The surface currents distribution at two frequencies 3.5 GHz and 5 GHz are shown in Figure 1(c-d). It is observed that the current density is significantly more near periphery instead of inner part of the circular patch. The current on the circular patch has two major components; vertical (along y-axis) and horizontal (along x-axis) which makes the

proposed UWB antenna to radiate linearly in two perpendicular directions. The measured and simulated field radiation patterns at 3.5 GHz and 5 GHz shown in Figure 1(c) and 2 respectively justify this argument.

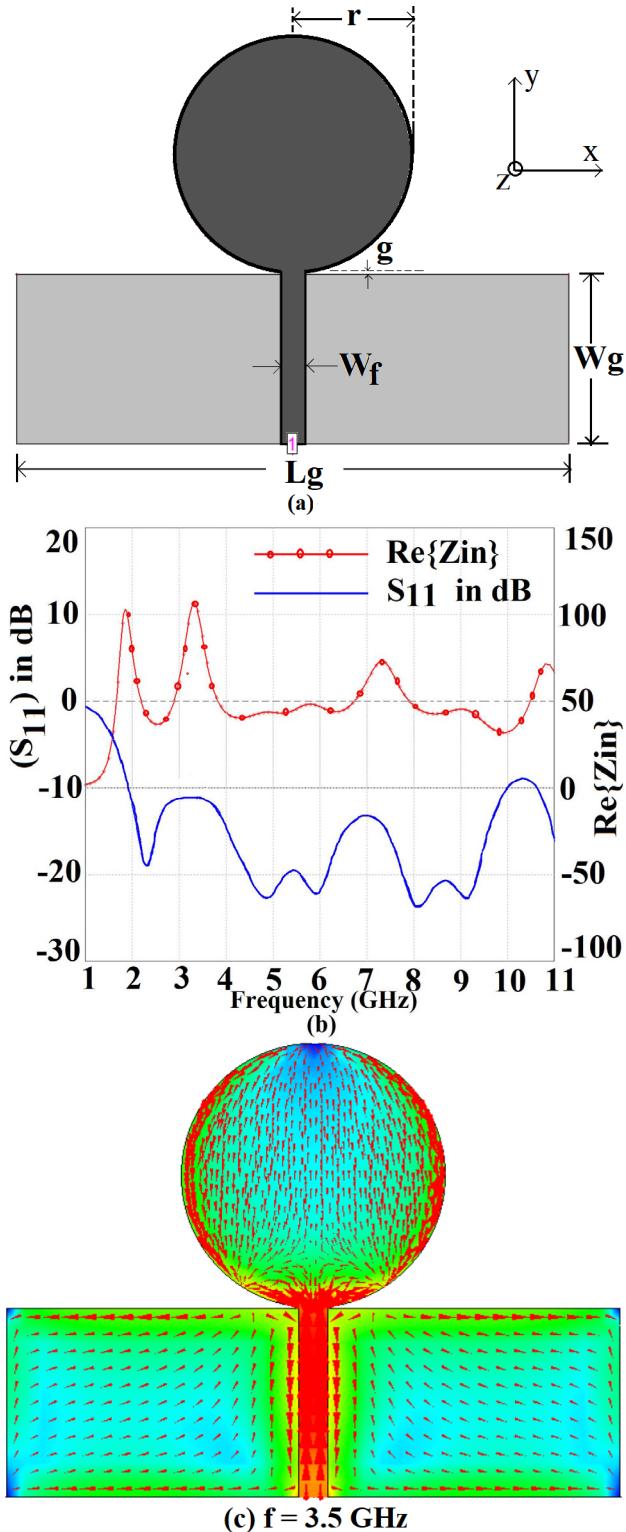


Fig. 1. (a) Circular shape UWB antenna, (b) Resonance plot and return loss (S_{11}) for circular shape UWB antenna, (c) surface current distribution at 3.5GHz.

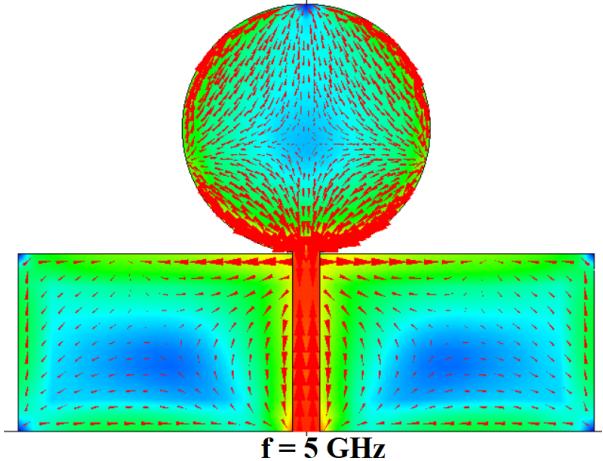


Fig. 2. Surface current distribution at 5GHz.

The field radiation pattern and gain of the antennas were measured using Rhode & Schwarz make spectrum analyzer (FSC-6) and RF source (SMB 100A) in the minimum reflection surroundings with more than $2D^2/\lambda$ between reference antenna and proposed UWB antenna under test. The radiation patterns of circular shape UWB antenna are shown in Figure 3 and 4 for two frequencies of 3.5 and 5 GHz respectively. The cross polar components are more since the currents in orthogonal directions are present in the radiating patch at above mentioned frequencies.

(—□—) E/H Co-polar Sim (—△—) E/H x-polar sim
 (—○—) E/H Co-polar meas (—●—) E/H x-polar meas

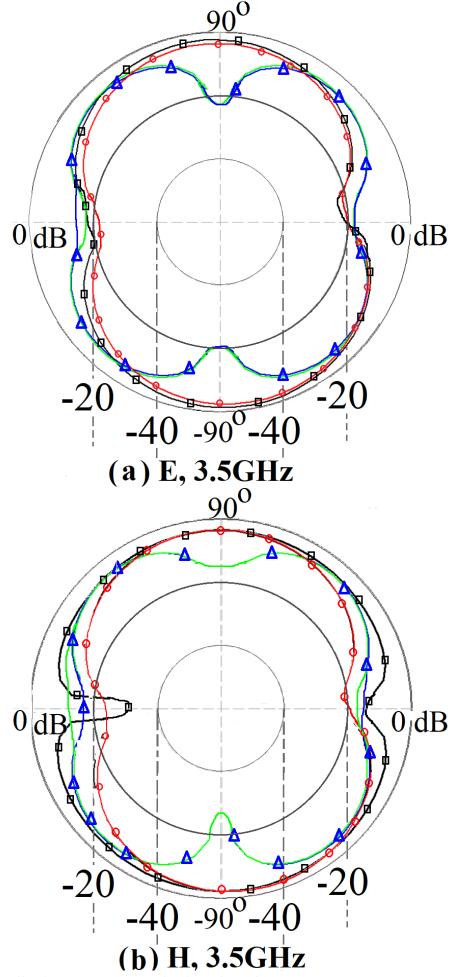


Fig. 3. Radiation patterns of Circular shape UWB antenna at $f = 3.5$ GHz.

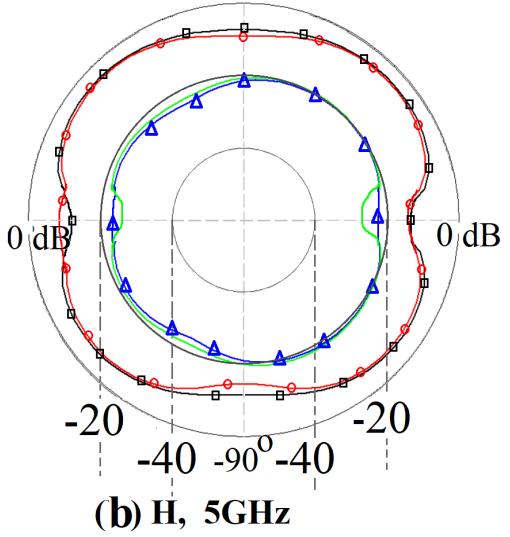
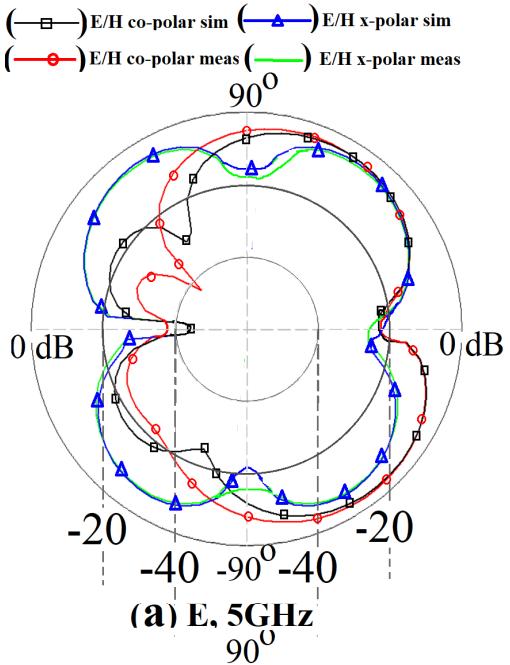


Fig. 4. Radiation patterns of Circular shape UWB antenna at $f = 5$ GHz.

III. NOTCH CHARACTERISTICS USING SLITS AND VARYING SLIT POSITION

Further, in this paper an innovative design of circular antenna for band notched application is proposed by adding pair of slits diametrically opposite and orthogonal to feed line. The notch characteristics in circular patch is obtained by these geometrical modifications which leads to change in the resonance frequencies and input impedance (Z_{in}) for the various resonant modes at various frequencies across UWB range. Either by altering length of the slits (L_s) or by altering the angular separation between slits (θ) the notch band can be tuned. Within the notch band, return loss (S_{11}) of more than -6.0 to -5.0 dB accounts for more than 45% of reflected power which results into minimization of an interference with other applications working in the same frequency range.

The circular antenna with pair of slits is shown in Figure 5. The pair of slits in circular patch changes the resonance frequencies and input impedance (Z_{in}) for the various resonant modes at respective frequencies which results into notch response across 3 – 10 GHz band.

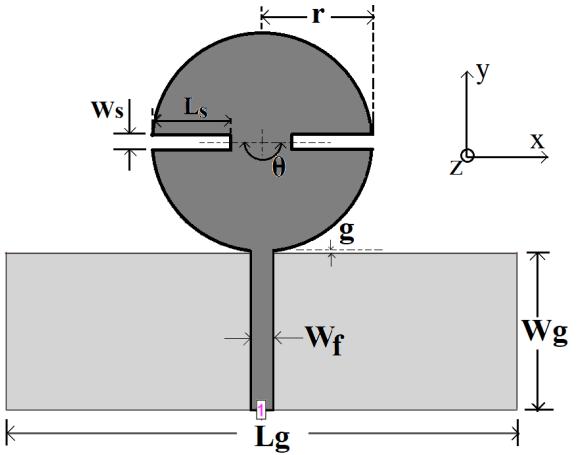


Fig. 5. Geometry of Slit Loaded Circular shape UWB antenna

The tuning of notch band can be obtained by varying the length of slits (L_s). The pair of slits in circular patch varies input impedance (Z_{in}) and the resonance frequencies of the various resonant modes at respective frequencies across UWB band results into notch characteristics. The length of the slot (L_s) tunes the notch band. As the slot length (L_s) increases the input impedance around 3.2 GHz gets increased above 250Ω while that around 3.4 GHz gets lowered below 20Ω which results in mismatch yielding a notch. For the proposed antenna the variations in input impedances for various length of slits (L_s) are shown in Figure 6(a). Also, for fixed value of slit length and width the varying angular separation (θ) alters the input impedance (Z_{in}) and the resonance frequencies of the various resonant modes at respective frequencies across UWB band results into notch characteristics. As seen from Figure 6(b) as the θ decreases from 200° to 140° , the input impedance of the TM_{21} mode increases beyond 200Ω for the frequencies around 3.2 GHz while that for frequency around 3.4 GHz decreases below 25Ω which results in mismatch yielding a notch.

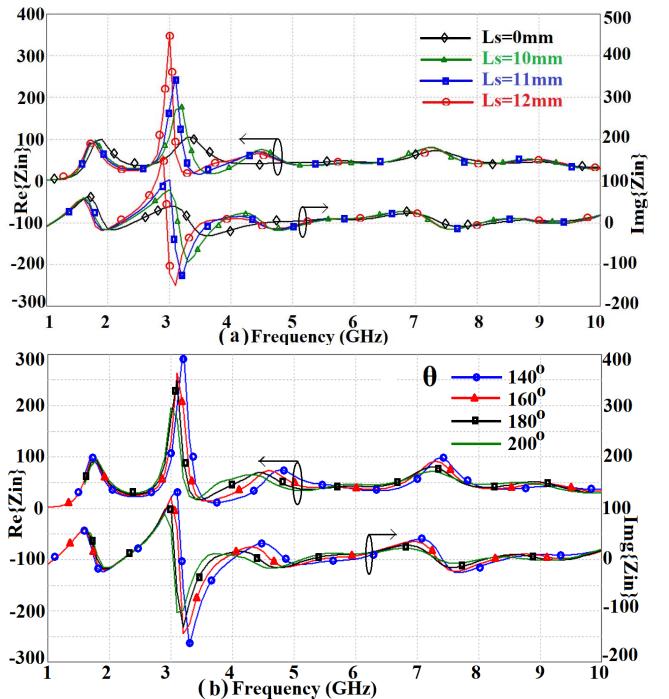


Fig. 6. Resonance plots for (a) for varying slit length (L_s) and (b) for varying angular separation between slits (θ) for fixed value of slit width $W_s=2\text{mm}$ for slit loaded circular antenna.

The variations in return loss for various length of slit (L_s) are shown in Figure 7. As seen from S_{11} plots for fixed value of width of slit $W_s = 2\text{mm}$, when slot length $L_s = 10\text{ mm}$, notch characteristic is obtained from 2.50 to 3.95 GHz. Further for stub length $L_s = 12\text{ mm}$, notch characteristics is realised from 2.2 to 3.5 GHz. Hence the variable slit length realizes tunable notch BW.

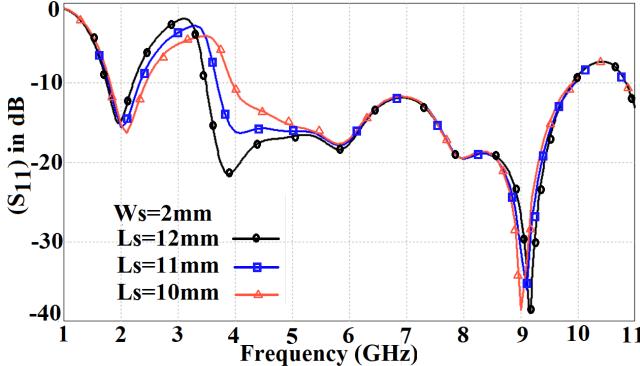


Fig. 7. Return loss plots for varying plots for varying slit length (L_s) for fixed value of $\theta=180^\circ$.

Further by varying the angular separation between the slits the tunable notch characteristics can be realized. The angular separation (θ) between the slits perturbs the current distribution of TM_{21} mode accordingly resulting into notch bands. The surface current distribution at 3.5 GHz for slit length $L_s = 11\text{ mm}$ and $\theta = 180^\circ, 160^\circ$ are shown in Figure 8(a-b).

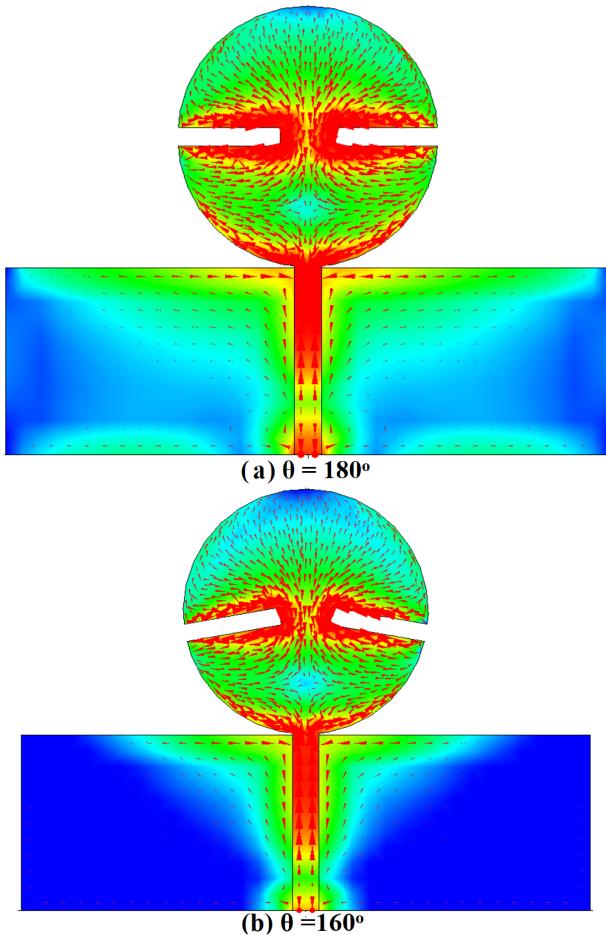


Fig. 8. surface current distribution of slit loaded circular shape antenna for (a) $\theta=180^\circ$, and (b) $\theta=160^\circ$.

As seen from Figure 9, for fixed value of slit width $W_s = 2\text{mm}$, the angular separation between slits $\theta = 120^\circ$, the notch band is realized in the frequency band 2.1- 4.82 GHz. While for $\theta = 220^\circ$, the notch band is obtained in the frequency range 2.5 - 3.5 GHz.

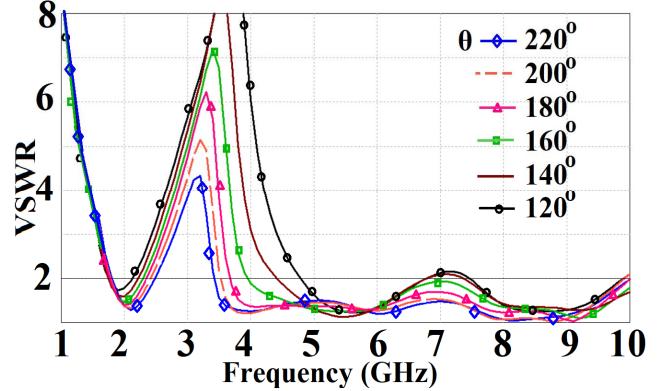


Fig. 9. VSWR plots for varying plots for varying angular separation θ for fixed slit length $L_s = 1\text{mm}$.

A pair of slit loaded circular antenna is fabricated for $L_s = 11\text{ mm}$ and angular separation $\theta = 160^\circ$. The measured and simulated S_{11} plots for the same are shown in Figure 10. The notch band frequency values obtained after simulation are, 2.28 to 3.92 GHz. The respective measured values are, 2.33 to 3.98 GHz respectively.

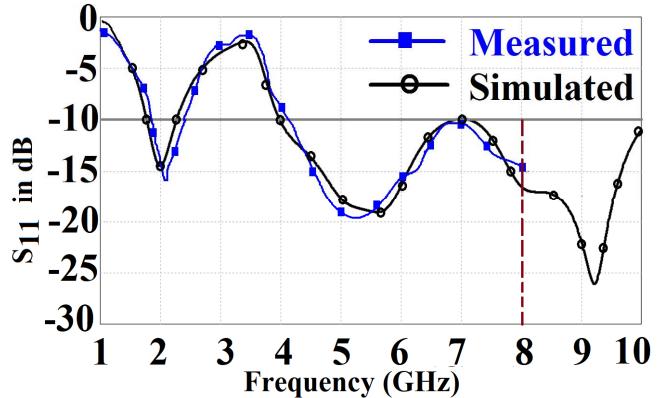


Fig. 10. Simulated and measured return loss plots for slit length $L_s = 11\text{mm}$, and $\theta = 160^\circ$.

Radiation patterns and gain are measured using standard horn antennas for the slit loaded circular antenna for $L_s = 11\text{ mm}$ and angular separation $\theta = 160^\circ$. The measured normalized radiation patterns at 5.5 GHz and 6.5 GHz in the principal planes are shown in Figure 11(a-d) while the measured gain up to 8 GHz of proposed antenna structure is shown in Figure 12. The antenna shows a stable omnidirectional radiation over UWB bands except in notched frequency bands. At higher frequencies, the gain deteriorates because the substrate becomes more lossy. Across the pass band, antenna with pair of slits having angular separation $\theta = 160^\circ$ shows radiation in broadside with gain of around 1–2 dBi. The gain of the antenna drops in the notch band of the antenna resulting into poor radiation.

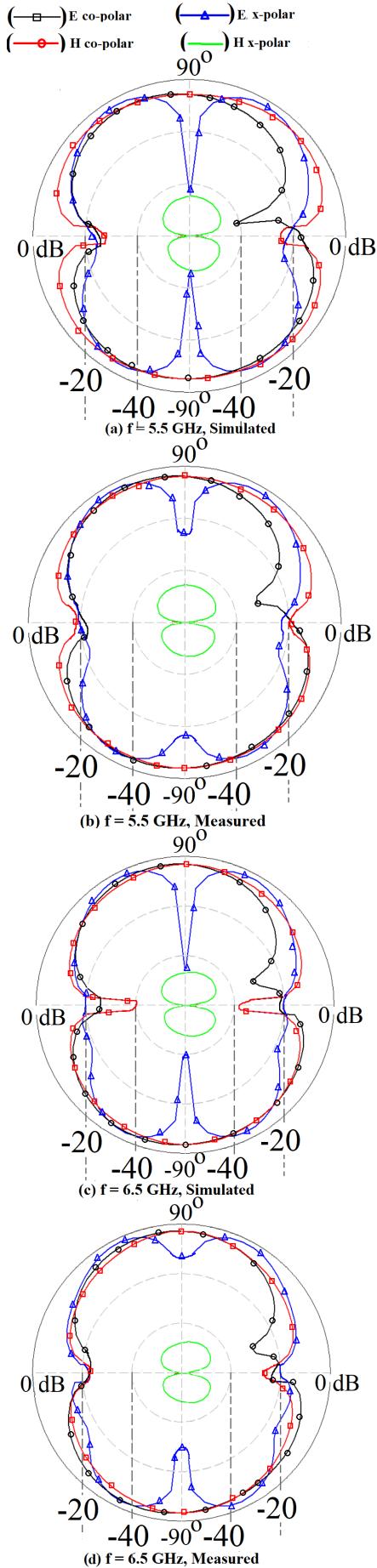


Fig. 11. (A-D) Simulated and measured radiation pattern for circular shape antenna with slit length $L_s = 11\text{mm}$, and $\theta = 160^\circ$.

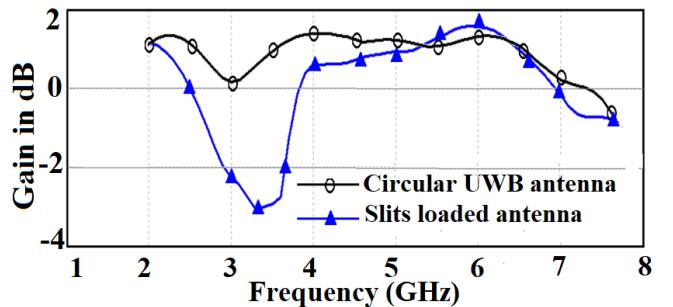


Fig. 12. Measured gain values for circular UWB antenna and slit loaded circular shape antenna.

The time domain performance of UWB applications is vital for pulsed systems. The group delay is a measure of the time delay of an impulse signal at various frequencies. For group delay evaluation, a pair of the proposed antenna configurations are placed at a distance of 0.4m face to face and aligned in the azimuthal plane at $\phi = 0^\circ$. Figure 13 shows little variation in the simulated group delay and magnitude of transfer function $|S_{21}|$ over the operating band except in notched frequency bands for the slit loaded circular antenna for $L_s = 11\text{ mm}$ and angular separation $\theta = 160^\circ$.

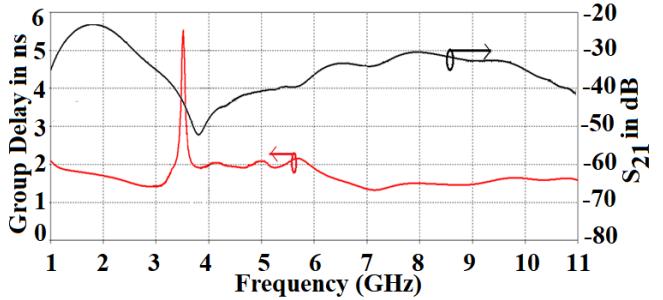


Fig. 13. Magnitude of transfer function $|S_{21}|$ and Group delay for slit loaded circular antenna.

The value of $|S_{21}|$ at higher frequency decreases because the substrate becomes more lossy at higher frequencies. The antenna structures can be optimized to minimize their inherent pulse spreading effect. To evaluate the pulse transmission characteristics of the proposed antennas, face to face orientation is implemented. The plot of applied Gaussian pulse to excite the transmitting antenna, and corresponding received pulse for face-to-face orientation are shown in Figure 14. The amplitude of the pulse is reduced and widened in the received pulse compared to applied pulse. The fabricated prototypes of configuration are shown in Figure 15(a-b).

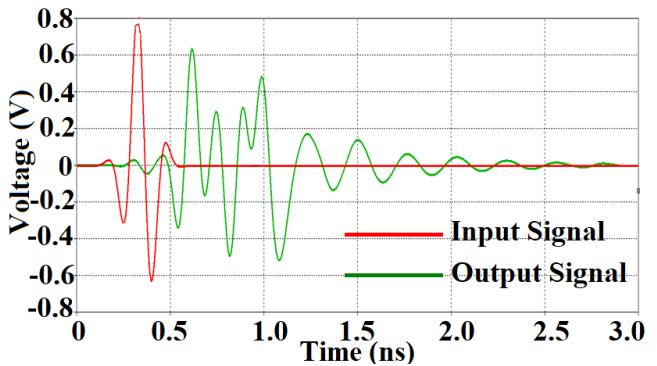


Fig. 14. Applied Gaussian pulse and output signal of slit loaded band notch antenna.

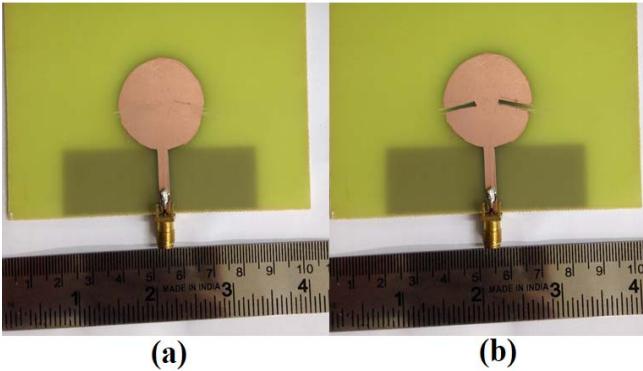


Fig. 15. Fabricated prototypes of (a) Circular shape UWB antenna, (b) slit loaded circular shape UWB antenna.

IV. CONCLUSIONS

The UWB circular shape antenna is proposed, which yields BW from 2 GHz to more than 10 GHz. The proposed design yields broadside radiation pattern. Further by altering UWB antenna design, tunable notch response is obtained either by using a pair of slits or by varying the angular separation between slits, over the UWB spectrum. The slit length or angular separation of slits modify the input impedance and resonance frequency at various resonant modes which results into notch characteristics in 2.2 – 5 GHz frequency band. The proposed structures show radiation in broadside with gain of around 1 – 2 dBi and little group delay variation over the pass band.

REFERENCES

- [1] Wong K L, Lin Y F, "Stripline-fed printed triangular monopole" Electron Letters. 1997; 33: pp. 1428–1429.
- [2] Ray K P, Thakur S S, Deshmukh R A, "UWB printed Sectoral monopole antenna with dual polarization", Microwave Optical Technology Letters; 54: 2066–2070, 2012.
- [3] Kumar G, Ray K P, "Broadband Microstrip Antennas", Artech House; 2002.
- [4] Ray K P , "Design aspect of printed monopole antenna for ultrawide band applications". International Journal Antennas Propagation; pp. 1–8, 2008.
- [5] Bahl I, Bhartia P, Ittipiboon A, and Garg R, "Microstrip Antenna Design Handbook", Artech House, 2001
- [6] Ray K P and Ranga Y, "Printed rectangular monopole antennas", IEEE Antenna Propagation Society International Symposium, 2006, pp. 1693–1696
- [7] Verbiest J R and Vandenbosch G A E, "Small-size planar triangular monopole antenna for UWB WBAN applications", Electronic Letters, 42, 2006, pp. 566–567.
- [8] Hong T, and Liu Y, "A novel monopole antenna for ultra-wide band application", Microwave and Optical Techology Letters Vol. 52, No. 12, December 2010, pp. 2694 – 2696.
- [9] Sanming H., Wenbin D., "A Balloon-In Shaped Monopole Antenna for Passive UWB-RFID Tag Applications", IEEE Antennas and Wireless Propagation Letters, vol. 7, 2008.
- [10] Liang J, Chen X, and Parini C, "Study of a printed circular disc monopole antenna for UWB systems", IEEE Transaction Antennas and Propagation., vol. 53, pp. 3500 –3505, 2005.
- [11] Deshmukh A A, Mohadikar P V, " Modified rectangular shape patch antennas for ultra-wide band and notch characteristics response", Microwave Optical Technology Letters, 2017;59:1524–1529.

Analysis and Resonant Length Formulation of Dual Band Microstrip Antenna with Modified Ground

Poonam Kadam

Research Scholar, EXTC Dept.
SVKM's D J Sanghvi College of
Engineering, Vileparle(west)
Mumbai, India
poonam.kadam@djsce.ac.in

Sanjay Deshmukh

Research Scholar, EXTC Dept.
SVKM's D J Sanghvi College of
Engineering, Vileparle(west)
Mumbai, India
sanjay.deshmukh@djsce.ac.in

Akshay Doshi

M.E., EXTC Dept.
SVKM's D J Sanghvi College of
Engineering, Vileparle(west)
Mumbai, India
doshiakshay4192@gmail.com

Amit A. Deshmukh

Prof. & Head, EXTC Dept.,
SVKM's D J Sanghvi College of
Engineering, Vileparle(west)
Mumbai, India
amit.deshmukh@djsce.ac.in

Abstract—A dual band microstrip antenna with defected ground plane is proposed in this paper. The dual bands are achieved by appropriately embedding pair of slots on the ground plane to tune TM_{30} mode resonant frequency. These slots essentially reduces TM_{30} mode frequency and places it closer to fundamental mode frequency. Broadside radiation pattern is observed at both the frequencies with maximum gain of nearly 3 dBi. The effect of the ground slot is studied and are portrayed by showing resonance plot for variation in the slot length. Empirical formulation of resonant length for fundamental as well as higher order TM_{30} mode frequency is suggested in the paper. The calculated frequencies obtained using the proposed formula closely matches with the simulated values. Thus these formulas can be applied to design dual band antenna with similar configuration at any given operating frequency.

Keywords—defected ground plane, dual band, resonant length formulation, microstrip antennas

I. INTRODUCTION

Compact Multiband antennas are generally preferred in realizing the modern wireless technology devices. Microstrip antennas being planar and conformal can be most suitably placed on the body of miniaturized wireless system. These antennas can easily be integrated with rest of the microwave components in the system [1]. Most of the wireless devices support multiple applications that demands different frequency spectrum such as IEEE 802.11a wireless band covers 5.15-5.725 GHz in Europe, IEEE 802.16a Wi-Max allows data transmission in the range 10-66 GHz and 2-11GHz for 802.11d. In order to accomplish multiple frequency demands for wireless communication devices multiband antennas are gaining more interest in recent years. Microstrip antennas though is known to be low profile antenna nevertheless has few inherent drawbacks such as low gain and small bandwidth [2]. Lots of techniques are discussed and reported in literature that highlights on achieving multiband response or improving the bandwidth of the microstrip antennas like gap coupled configuration [3, 4] where the bandwidth enhancement results due to coupling between the fed patch and the parasitic patches. When the coupling between the two modes results in formation of loop

inside the VSWR 2 circle in the smith chart it yields broadband response. For multiband response the loop size is large so loop is not seen inside the VSWR 2 circle rather the resonant modes lie within this circle. Other techniques includes adding stubs, stacked structures etc. All these methods increase the antenna size. Another solution which is most widely used to realize multiband or for enhancing bandwidth is embedding slots on the radiating patch [5-10]. When the slots are placed at the appropriate position on the patch it also results in antennas size reduction [11-13]. Other methods reported in literature to achieve compactness includes shorted plane, shorted pins and edge shorted configuration. Recently huge amount of investigations are taking place on the defected ground plane antennas to reveal their benefits in improvising antenna characteristics. It has been projected [11] that slots on the ground plane of the antenna has similar effects on resonance frequencies as that of antennas with slotted patch. The ground plane provides the return path to the signal and therefore defect on the ground plane perturbs its current distribution which in turn disturbs the current distribution on the radiating patch. The papers in literature has shown bandwidth, gain and size improvement over conventional microstrip antennas by utilizing DGS [12-19]. However most of the papers simply claim the results with no clear and detailed explanation about the reason for obtaining such results. Also systematic guidelines and resonant length formulation to achieve multiband response in rectangular microstrip antenna (RMSA) with defected ground structure is not reported in literature. This paper aims to provide the proper guidelines for realizing ground slotted antennas for improved characteristics.

In this paper a dual band rectangular microstrip antenna is realized by incorporating two slots on the ground plane under the patch close and parallel to the radiating edges. These slots plays an important role in bringing TM_{30} higher order mode frequency closer to the fundamental TM_{10} mode frequency. The two bands obtained has centre frequency at 712 MHz and 1600 MHz respectively. The radiation pattern is broadside at both the frequencies with cross polar levels less than 20 dBi. The maximum gain obtained at the broadside direction is around 3 dBi. The effect of the

position and length of these slots are studied in this paper by investigating the shift in resonant mode frequencies due to these slots. Also the resonant length formulas are proposed in this paper which can be applied to design the DGS antenna to work at any desired frequency. The proposed antenna was initially simulated using IE3D software and was later verified by performing experimental measurement. The antenna is designed with Taconic substrate having dielectric constant of 3 and loss tangent 0.002. Coaxial feeding technique is utilized to feed the antenna. Feeding is done using 50 ohm SMA connector having inner wire radius of 0.06 cm. The radiation plot and the impedance measurement were carried out using VNA and ZHV - 8.

II. RMSA WITH PAIR OF SLOTS ON THE GROUND PLANE

The RMSA without any slots shown in Fig. 1(a) is studied first in order to observe the excited modes for the given feed point location (-15, 0). Since the feed is placed at the center of the width along the patch length thus TM₀₁ mode is not getting excited. The dimensions of RMSA is 10 cm x 6 cm is backed by ground plane of size 12 cm x 8 cm, such that the MSA resonates at 860 MHz in TM₁₀ mode. The fundamental and higher order mode frequencies for the given patch calculated using equation (1) are 860 MHz for TM₁₀ mode, 1730 MHz for TM₂₀ mode, 2590 MHz for TM₃₀ mode and 2880 MHz for TM₀₂ mode. The patch is simulated using IE3D simulation software which shows resonance peaks at the same frequencies. The resonance plot of this antenna displays first four resonant peaks corresponding to TM₁₀, TM₂₀, TM₃₀ and TM₀₂ modes. This is claimed by observing the surface current distribution at each of these modes

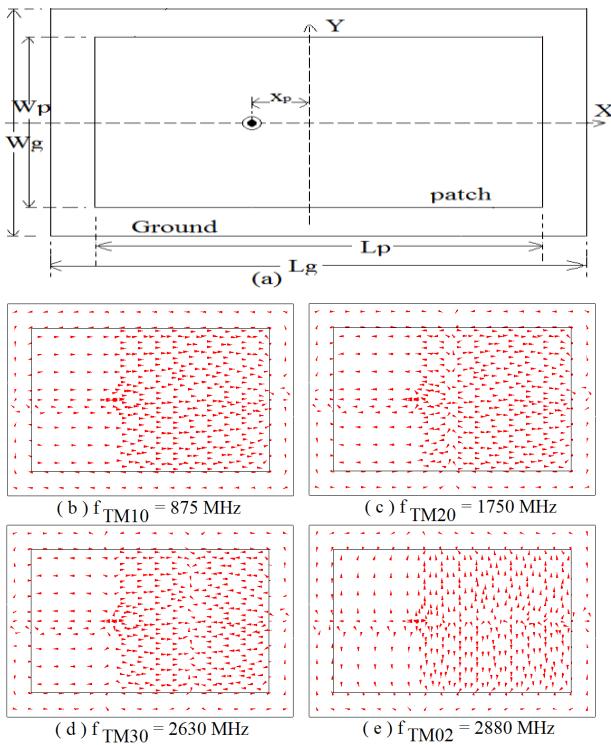


Fig. 1(a) RMSA with feed point at $x_p = -15$ mm (b - e) Current distribution at resonant peaks

In order to realize multiband antenna it is necessary that the resonant modes giving rise to multiband response must lie within VSWR 2 circle in the smith chart and they must exhibit similar radiation pattern. The radiation pattern at TM₁₀ and TM₃₀ mode are in broadside direction as shown in Fig. 2(a, b) whereas TM₂₀ and TM₀₂ modes yields conical pattern. From the resonance plot it is clearly seen that the TM₁₀ mode and TM₃₀ mode frequencies are far apart. The fundamental mode frequency is at 875 MHz whereas TM₃₀ mode frequency is at 2630 MHz Thus the objective of this work is to bring the higher order TM₃₀ mode closer to the fundamental mode.

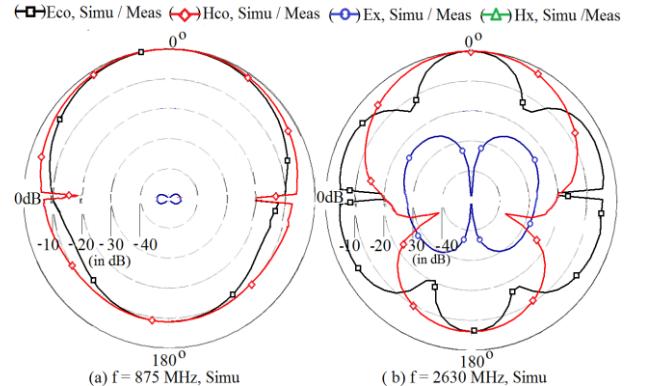


Fig. 2(a, b) Simulated Radiation pattern of RMSA without slots

Two slots are etched on the ground plane close and parallel to the radiating edges of the patch so as to perturb the surface current distribution at TM₃₀ mode frequency. This slot position is chosen as it is the maximum current position for TM₃₀ mode. The height of the slot is increased so that they are orthogonal to the current distribution at this mode. When the currents encounter slots orthogonal to it, the current path is elongated and since the resonant length at this mode is increased the frequency at this mode reduces. The longer the slot, the more will be reduction in its frequency. These slots has very less effect on the fundamental mode frequency. The effect of variation in the slot length on the fundamental and TM₃₀ mode frequencies is analyzed and is depicted in the resonance plot shown in Fig 4(a, b). The impedance at TM₁₀ is observed to be very small. In order to achieve impedance matching the feed point is shifted little towards the radiating edge.

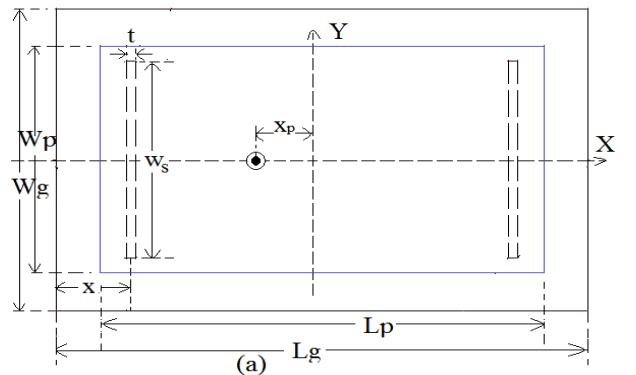


Fig. 3(a) DGS RMSA with feed point at $x_p = -15$ mm, Front view (b) back view

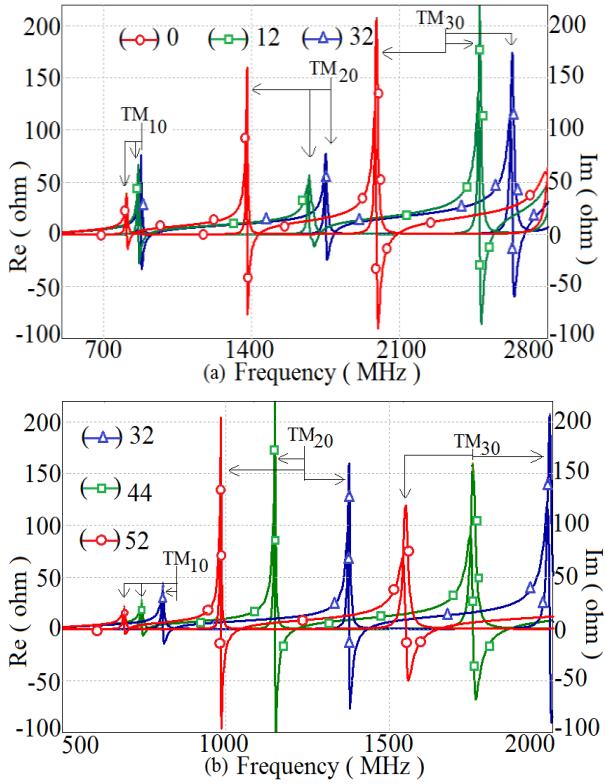


Fig. 4(a, b) Resonance plot for varying vertical slot length w_s (unit in mm)

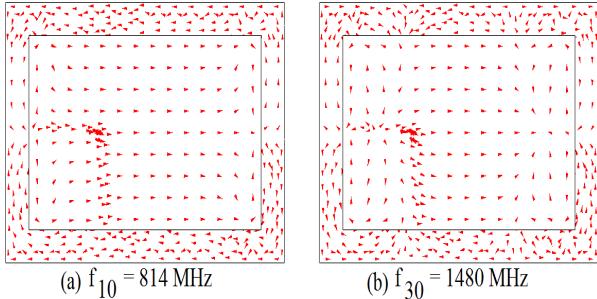


Fig. 5 (a, b) Current distribution at TM_{10} and TM_{30} modes frequencies vertical slot length $w_s = 48\text{mm}$

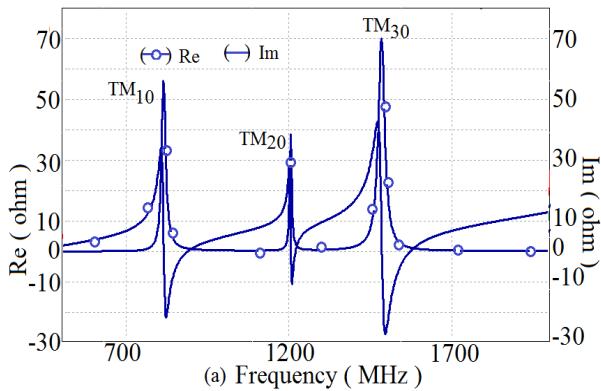


Fig. 6 Resonance plot at vertical slot length $w_s = 52\text{ mm}$ and feed point $(-23,0)$

The current distribution and the resonance curve for vertical slot length of 48mm is shown in Fig. 5(a, b) and Fig. 6. Fig.7 shows the impedance plot for the proposed antenna. The gain plot as well as simulated and measured radiation pattern of the optimized antenna at the dual frequencies is shown in Fig. 8, Fig. 9(a, b) and Fig. 10(a, b). As can be seen in Fig.8 the gain of the antenna at TM_{10} and TM_{30} mode are around 3dBi and 1.5 dBi respectively. The lower gain is primarily due to the finite ground plane. The slot cuts on the ground also reduces the gain because of back radiations but its effect is less significant here as the slot dimension is considerably small. As shown in the Fig. 9(a, b) and Fig. 10(a, b) the radiation pattern at the two frequencies are in broadside directions with E and H planes aligned along $\Phi = 0^\circ$ and 90° respectively. The deviation in the measured radiation pattern from the simulated radiation pattern observed because the pattern measurement for the proposed antenna was carried out in the environment with minimum reflection which means that the surrounding objects cannot absorb electromagnetic waves completely.

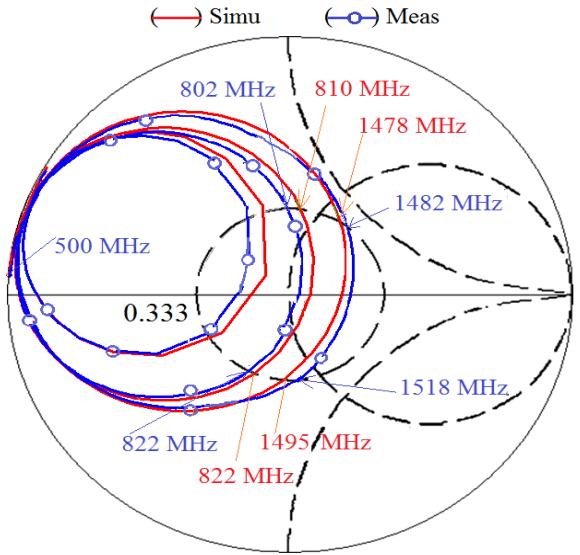


Fig. 7 Smith chart plot for optimized RMSA antenna with modified ground plane

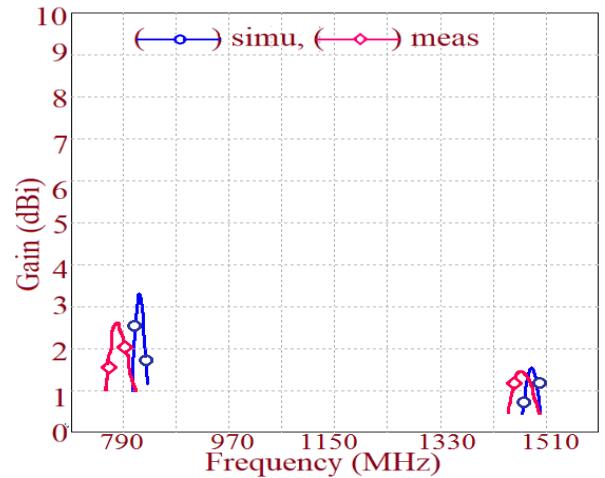


Fig. 8 Gain of optimized RMSA with slots on ground plane

The dual frequency response of the antenna is verified through experiment and the image of the fabricated prototype is shown in Fig. 11(a, b). The simulated and measured frequencies of proposed antennas are $f_{10} = 812$ MHz and $f_{30} = 1600$ MHz and $f_{10} = 714$ MHz and $f_{30} = 1690$ MHz respectively. The bandwidth obtained at these bands are approximately 15 MHz and 20 MHz respectively as shown in smith chart in Fig. 7.

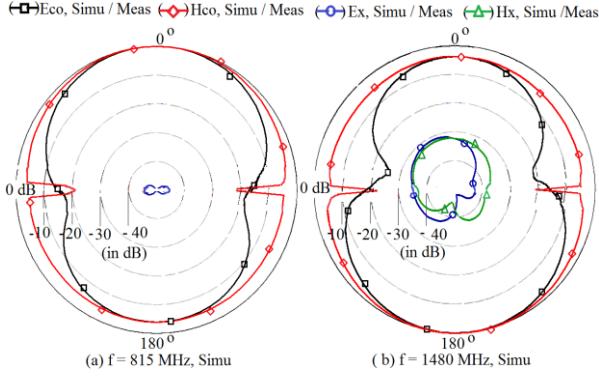


Fig. 9(a, b) Simulated Radiation pattern of RMSA with pair of slots on ground plane

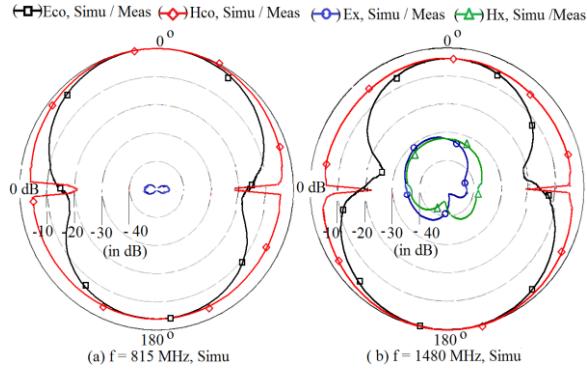


Fig. 10(a, b) Measured Radiation pattern of RMSA with pair of slots on ground plane

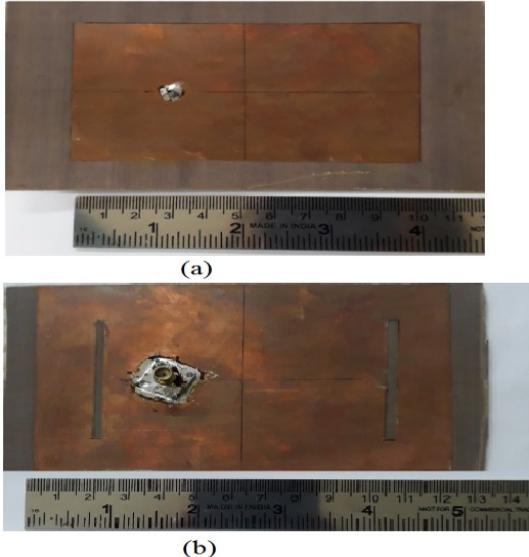


Fig. 11(a) Fabricated prototype of DGS RMSA with feed point at $x_p = -23$ mm, Front view (b) back view

III. RESONANT LENGTH FORMULATION

The resonant length formulation is done for both the dual frequency resonant modes. The formula proposed is verified by comparing the simulated resonant frequencies with the calculated frequencies for different slot lengths. The formulations were made by observing the surface current distribution at these modes. The current distribution on patch at TM_{10} mode is along the length of the patch and shows half wavelength sinusoidal variation. The vertical slots on the ground plane are orthogonal to the current distribution and thus forces the current path of the patch to change. The amount of current perturbation on the patch is function of the location and the length of the slot. When the slots are placed at the maximum current location ie at the center it will deviate the distribution to its maximum. Whereas when the slots are moved towards the radiating edges the current perturbation is reduces. Therefore sine term is included in the formula given in eq. (1). The current perturbation is also function of the slot length. For very small slots the currents are not perturbed and the current is along the patch length. The amount of perturbation increases and follows the slot as the slot length is increased as shown in Fig. 12. This is accounted by the $A^*(w_s / W)$ term. The Fringing fields at the edges of the antenna is accounted by adding $2\Delta L$ term in the resonant length formula.

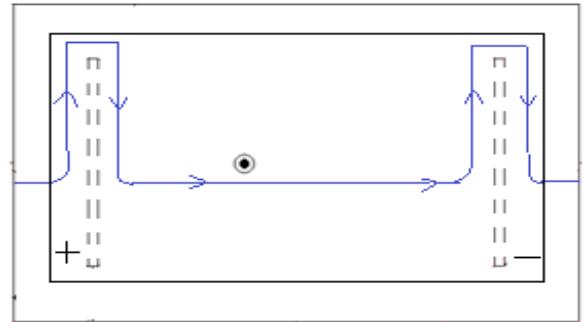


Fig. 12 Current distribution at TM_{10} mode due to pair vertical slots on the ground plane

$$L_e = L + A \times (2w_s) \times \left(\frac{w_s}{W_p} \right) \sin(\pi x/L) + 2\Delta L \quad \dots \dots \dots (1)$$

Where $A=0.75$;

$$\Delta L = \frac{h}{\sqrt{\epsilon_{re}}}$$

The resonance frequencies at TM_{01} mode were calculated using eq. (2)

$$f = \frac{1}{2\sqrt{\epsilon_{re}}} \sqrt{\left(\frac{m}{L_e} \right)^2 + \left(\frac{n}{W_e} \right)^2} \quad \dots \dots \dots (2)$$

The calculated frequencies for varying slot lengths are computed using the proposed formula and is compared with the simulated frequencies and the percentage error is calculated using equation (3).

$$E = \frac{f_{\text{cal}} - f_{\text{sim}}}{f_{\text{sim}}} \quad \dots \dots \dots \quad (3)$$

The percentage error and resonance frequencies for the calculated and simulated values for varying slot lengths is observed and shown in Fig. 13. It clearly shows that the percentage error is less than 5% of error.

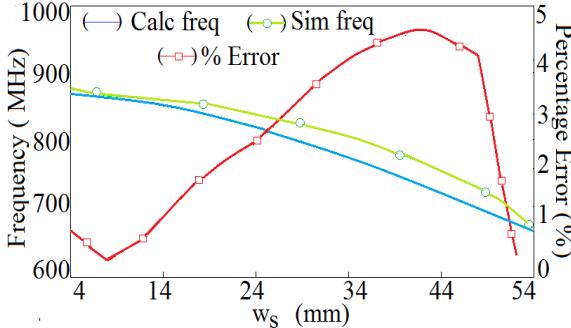


Fig.13 Percentage Error % and simulated versus calculated frequency comparision at TM₁₀ mode

For TM₃₀ mode three half wavelength variations are seen by the patch length. The resonant length formula for TM₃₀ mode is given in equation (2). The current path at TM₃₀ is shown in Fig. 14.

$$L_E = L + 2A \times w_s \times \left(\frac{w_s}{W_p} \right) \sin(3\pi x/L) + 2\Delta L \quad \dots \dots \dots \quad (2)$$

Where, A=0.77;

$$\Delta L = \frac{h}{\sqrt{\epsilon_r e}}$$

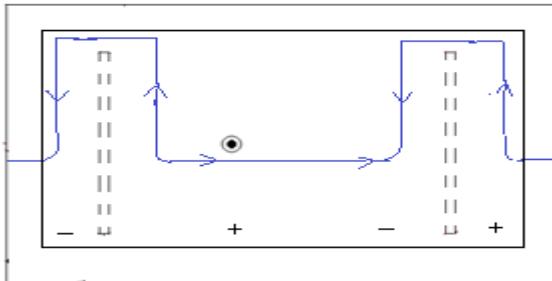


Fig.14 Current distribution at TM₃₀ mode due to pair vertical slots on the ground plane

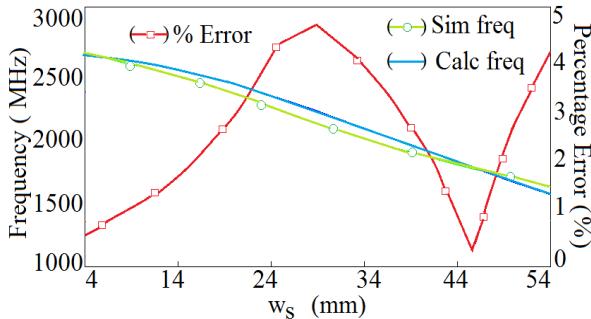


Fig.15 Percentage Error % and simulated versus calculated frequency comparision at TM₃₀ mode

The percentage error and resonance frequencies for the calculated and simulated values for varying slot lengths is observed and shown in Fig. 15. It clearly shows that the percentage error is less than 5% of error.

IV. CONCLUSIONS

Dual band microstrip antenna with defected ground plane is proposed in this paper. Detailed modal analysis of the antenna is demonstrated in the paper. Further it proposes the resonant length formulation for TM₁₀ and TM₃₀ mode frequencies of the antenna. The Proposed antennas are simulated and verified with fabricated and measured results. The simulated and measured results matches very closely. The resonant length formula proposed at modified patch modes can be applied to realize similar antenna at any desired frequency based on the target application. The frequencies calculated using these formulations matches well with the simulated values with less than 5% error.

REFERENCES

- [1] G. Kumar, K. P. Ray, "Broadband Microstrip Antenna", Artech House, Boston, 2003.
- [2] K. L. Wong, "Compact and Broadband Microstrip Antennas", John Wiley & sons, Inc., New York, USA, 2002.
- [3] C. K. Anandan, P. K. Mohanan and K. G. Nair, "Broad band gap coupled Microstrip antenna", IEEE Transaction on Antennas and Propagation, pp. 1581 – 1586, 1990.
- [4] A. Deshmukh, A. Joshi, Tirolkar, "Broadband slot cut gap coupled proximity fed E-Shaped microstrip antenna", International Journal of Computer Applications, pp. 15 – 18, 2011.
- [5] K. L. Wong and K. Ping "Compact dual-frequency microstrip antenna with a pair of bent slots, Electronics Letters, Vol. 34, 1998.
- [6] H. Elftouh, N. Touhami, M. Aghoutane, S. Amrani, A. Tazon, and M. Boussouis, "Miniaturized Microstrip Patch Antenna with Defected Ground Structure", Progress In Electromagnetic Research, Vol.55, 2014.
- [7] T. Wang, Y. Z. Yin, J. Yang, Y. L. Zang and J. J. Xie, "Compact triple band antenna using defected ground structure for WLAN/WiMAX applications", Progress In Electromagnetic Research Letters, pp. 155 – 164, 2012.
- [8] M. Chakraborty, B. Rana, P. Sarkar & A. Das, "Design and Analysis of a compact rectangular microstrip antenna with slots using Defective ground structure", Elsevier, 2012.
- [9] G. F. Khodaei, J. Nourima and C. Ghobadi, " A Practical Miniaturized U-Slot Patch Antenna with Enhanced Bandwidth", PIER-B, volume 3, pp. 47-62, 2008.
- [10] J. Ghalibafan and A. R. Attari, " A new dual band Microstrip Antenna with U-Shaped slot", Progress In Electromagnetic Research B, pp. 221 -235, 2010.
- [11] A. A. Deshmukh and G. Kumar, " Compact Broadband U-Slot loaded rectangular Microstrip Antennas", Microwave and Optical Technology Letters, volume 46, 2005.
- [12] K. Weli, J. Y. Li and Wang, "A ne technique to design circularly polarized microstrip antenna by fractal defected ground structure", IEEE Transactions on Antenna and Propagation, volume 65, 2017.
- [13] M. Khandewal, B. Kumar, S. Dwari, S. Kumar and A. K. Gautam, " Compact stacked microstrip patch antenna with defected ground structure for WLAN/ WiMax applications", AEU- International Journal of Electronics and Communications, volume 69, 2015.
- [14] H. V. Prabhakar, U. K. Kummuri, R. M. Yadahalli and V. Munnappa, "Effect of various meandering slots in rectangular microstrip Antenna ground plane for compact broadband operation", Electronics Letters, volume 43, 2007.

- [15] J. Kuo1 and K. L. Wong, "A compact microstrip antenna with meandering slots on the ground plane", *Microwave and optical Technology Letters*, volume 29, 2001.
- [16] W. C. Liu, , C. M. Wu, and Y. Dai, "Design of Tripple.Frequency Microstrip fed Monopole antenna using Defected Ground Structure", *IEEE transactions on antennas and propagation*, volume 59, 2011.
- [17] B. Kumar, M. Kumar and A. Kumar," Analysis and Design of compact high gain microstrip antenna with defected ground structure for wireless applications", *Wireless personal communications*, November 2016.
- [18] L. Li, S. Rao and B. Tang," A novel compact dual band monopole antenna using defected ground structure", *International workshop on Microwave and millimetre wave circuits and system Technology*, 2013.
- [19] W. Chung Liu, C. M. Wu and Y. Dai, " Design of triple frequency microstrip fed monopole antenna using defected ground structure", *IEEE Transactions on Antenna and propagation*, volume 59, 2011.

Proximity Fed Broadband Equilateral Triangular Microstrip Antenna Using Parasitic Rectangular Patches

Sanjay Deshmukh
Research Scholar, EXTC Dept
SVKM's D J Sanghvi College of
Engineering, Vile Parle(west)
Mumbai, India
sanjay.deshmukh@djsce.ac.in

Akshay Doshi
M.E., EXTC Dept.
SVKM's D J Sanghvi College of
Engineering, Vile Parle(west)
Mumbai, India
doshiakshay4192@gmail.com

Amit A. Deshmukh
line 2: Prof. & Head, EXTC Dept.
SVKM's D J Sanghvi College of
Engineering, Vile Parle (west)
Mumbai, India
amit.deshmukh@djsce.ac.in

Abstract—Gap coupled design of proximity fed equilateral triangular microstrip antenna along with variations of rectangular microstrip antennas are proposed. Design of equilateral triangular microstrip antenna provides a bandwidth of 302 MHz (25%) with broadside gain of 7.6 dBi. Enhancement in bandwidth and gain is realized by gap coupling rectangular patches. Optimum result with VSWR bandwidth of more than 670 MHz (57%), showing peak gain of 9.7 dBi is obtained in gap coupled configuration with two adjacent layers of two and four rectangular patches. Proposed design is simpler in implementation and measured results show close agreement.

Keywords—gap coupling, proximity fed, rectangular patches, peak gain

I. INTRODUCTION

With the rapid development of modern wireless communication technology , the microstrip antenna is used for applications such as radars, satellite, Radio frequency identifications (RFIDs), telemetry, aerospace etc. In its conventional form microstrip antennas have certain limitations like narrow bandwidth, polarization impurity for better performance. Preferably , the microstrip antennas are required to have high gain, compact size, broad bandwidth, flexibility and easy to fabricate. The need of almost all the applications of the antenna are higher gain and larger bandwidth. Several techniques have been reported to increase bandwidth of microstrip antennas. This includes techniques such as usage of thicker substrate, low dielectric constant substrate, implementing various impedance matching and feeding techniques, using slots incorporated on MSA [1-2]. Parabolic reflector has high gain with the drawback of large size, non-planar, bulky structure restricting its use in mobile communication. Microstrip antennas (MSA) have advantage of ease in integration with microwave integration circuits, easy to fabricate, less in cost and with planar structure [3].Gain can be increased by using line fed MSA arrays with the drawback of complexity in feeding structure, high cross polarization and low efficiency. Low efficiency in line feed MSA is due to losses and radiation in the feed network[4-6]. MSA reflect array eliminates requirement of line fed network, with the limitation of aperture blockage that reduces gain [7]. Space fed arrays were designed to remove drawback of aperture

blockage [8-9]. But as size of space fed MSA arrays increases, unequal excitation of space fed arrays elements results in small incremental gain . Gain can also be enhanced using Fabry-Perot cavity (FPC) resonator concept. FPC has a partially reflecting surface (PRS) and a ground plane. Metallic MSA array fed FPC is used to enhance the gain of antenna.But yields narrow bandwidth and has to be fabricated mechanically [10].In this paper, design of a microstrip bi Yagi and microstrip quad Yagi array antenna is proposed. The bandwidth of 8.1%, 7.1% and 5% for microstrip Yagi, bi Yagi, quad Yagi MSA is reported. Small bandwidth of quad Yagi due to shift of lower resonance of driven element to higher frequency is observed [11-12]. The Yagi antenna is bulky and needs complex feed design. The stacked configuration of multilayer Yagi array antennas are proposed to enhance the bandwidth by 16.6 %. But these configurations require complex design and are bulky [13-14].

In this paper, gap coupled variations of equilateral triangular microstrip antenna (ETMSA) with rectangular microstrip antenna (RMSA) are proposed. With variation in structure such as single ETMSA, ETMSA gap coupled with two RMSA, ETMSA gap coupled with four RMSA, ETMSA gap coupled with five RMSA and ETMSA gap coupled with six RMSA are discussed. Proximity fed single ETMSA yields BW of more than 302 MHz and gain more than 7 dBi. In the ETMSA with two RMSA gap coupled configuration BW of more than 550 MHz (49%) and peak gain of 8.4 dBi is obtained. Gain and bandwidth is enhanced by adding two more RMSAs which yields BW of more than 630 MHz (54%) and peak gain of 9.53 dBi. Further BW and gain is enhanced by adding three more RMSAs next to two RMSAs that provides BW of more than 665MHz (56%) and peak gain of 9.9 dBi. Also BW and gain is enhanced by adding four more RMSAs next to two RMSAs that provides BW of more than 670 MHz (57%) and peak gain of 9.7 dBi.

The above proposed broadband MSAs have been optimized using IE3D software and simulated results were verified with experimentation [15]. The proposed gap coupled broadband MSAs are optimized in 900–1800 MHz range. Simulated gain 9.7dBi was obtained for the proposed configuration. This frequency range is selected for proposed MSA as the antennas have narrower BW in the lower frequency range. For ETMSA gap coupled with RMSAs,

the detailed study is carried out by varying the length of RMSA and the gap between ETMSA and RMSA. MSA parameters were measured- input impedance, radiation pattern in the antenna lab using instruments like ZVH-8, SMB100A and FSC-6 by using ground plane of size 600mm X 600mm.

II. BROADBAND PROXIMITY FED ETMSA GAP COUPLED VARIATIONS WITH RMSA

A. Proximity fed ETMSA

In this basic configuration proximity feed is used to excite ETMSA as shown in Fig 1(a,b). Here the fundamental mode TM10 frequency of 1200 MHz is selected and ETMSA is optimized using a total substrate thickness of $0.1\lambda_0$. Radiating equilateral triangular patch is etched on glass epoxy substrate of thickness $h=1.6$ mm, dielectric constant=4.3 and loss tangent $\delta = 0.02$. Substrate is suspended above the ground plane with air gap ' h_a ' at a height of 24 mm. Proximity strip is placed at a height ' h_s ', which is 22 mm above the ground plane. For the above mentioned substrate dimensions and operating frequency, side length 'S' of ETMSA is found to be 108mm and horizontal length 'he' 94mm. By tuning dimension of proximity strip and its position below ETMSA, wider bandwidth is achieved. The simulated bandwidth is 302 MHz (25%).

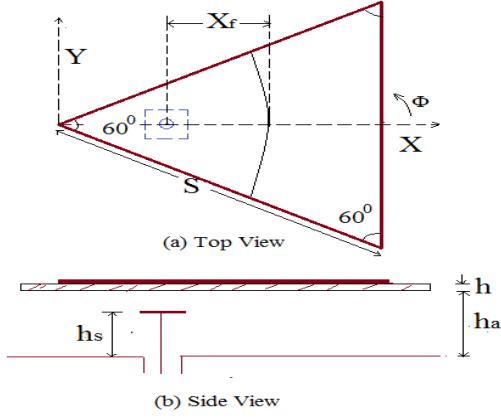


Fig. 1 (a, b) Proximity fed ETMSA

B. Proximity fed ETMSA gap coupled with two RMSA

This proposed structure with ETMSA gap coupled with two RMSA to enhance BW as shown in Fig. 2.

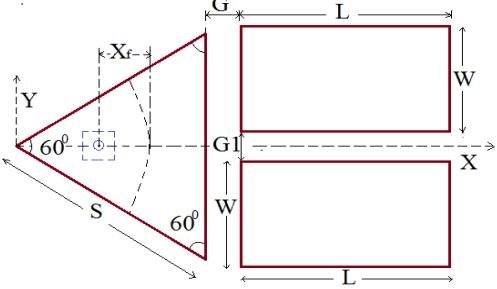


Fig.2. Proximity fed ETMSA gap coupled with two RMSA

ETMSA is coupled with two rectangular patches with length L and width W . The Gap between equilateral triangular patch and rectangular patches is G which decides

coupling in the structure. Gap between rectangular patches, $G1$ couples the two rectangular patches to enhance BW further. The gap between two RMSA is 5mm. Resonant frequency of parasitic rectangular patches are higher than driven patches frequency Parametric study was performed for the above configuration by varying L , X_f and G . Fig. 3 shows resonance curve for ETMSA and ETMSA gap coupled with two RMSA configurations. Two resonant peaks are observed at frequency $f_1=1000$ MHz and $f_2=1320$ MHz for ETMSA gap coupled with two RMSA configuration. Single resonant peak at frequency 1200 MHz is observed in ETMSA configuration.

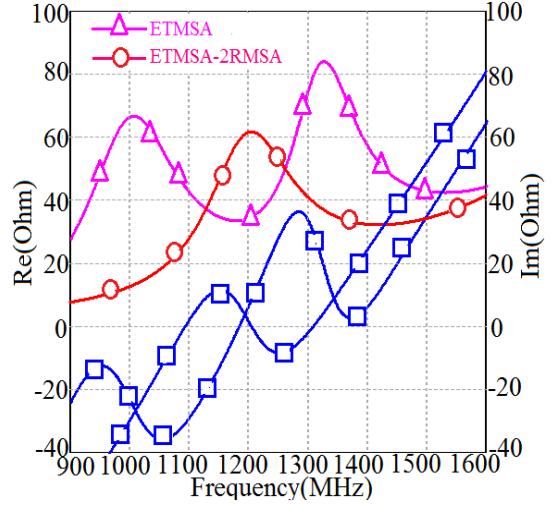


Fig.3. Resonance curve for ETMSA and ETMSA gap coupled with two RMSA

TABLE I. Parametric study of ETMSA gap coupled with two RMSA configuration

Sr.No	Length L in mm	Gap G in mm	Feed Xf in mm	BW in MHz	Gain in dBi
1	92	14	16	265	8.16
2	92	14	18	468	8.41
3	92	12	16	447	8.11
4	92	12	18	483	8.30
5	92	12	20	515	8.57
6	92	10	20	370	8.47
7	90	12	18	474	8.37
8	90	12	20	519	8.63
9	90	12	22	581	8.72
10	90	10	18	483	8.37
11	90	10	20	532	8.51
12	88	12	20	513	8.72
13	88	12	22	552	8.7
14	88	10	20	527	8.7
15	88	10	22	568	8.81
16	88	8	20	539	8.7
17	88	8	22	564	8.76

From table I, as length L of RMSA decreases, BW increases due to increased spacing between resonant frequency of ETMSA and RMSA. BW increases as gap G between ETMSA and RMSA decreases due to the increase in coupling between ETMSA and RMSA. Optimum BW and Gain obtained for parameter values of length $L=90$ mm, gap $G=12$ mm, feed $X_f=22$ mm. Fig.4. (a) and (b) shows ETMSA resonates at frequency 1000 MHz and surface current is unidirectional. Whereas in Fig.4.(b) RMSA resonates at frequency 1320 MHz and surface current is unidirectional as well.

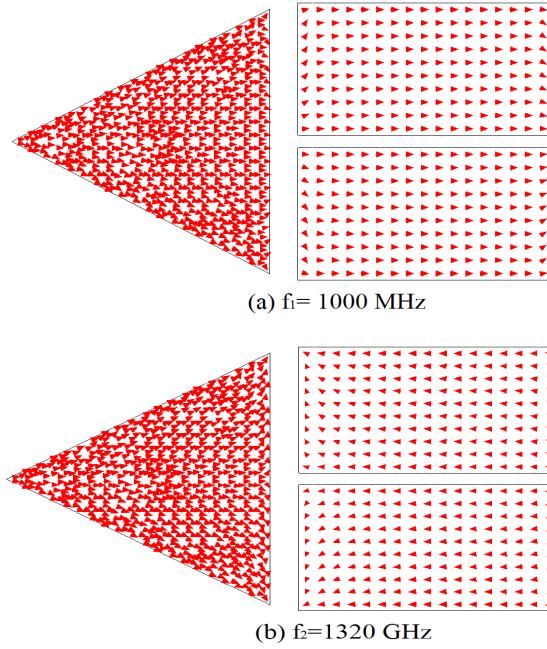


Fig.4. Surface current distribution for ETMSA gap coupled with two RMSA
(a) $f_1=1000$ MHz (b) $f_2=1320$ MHz

C. Proximity fed ETMSA gap coupled with four RMSA

This proposed configuration is ETMSA gap coupled with two RMSA further followed by two more RMSA. Gap between layers of RMSAs is G_2 which is 8 mm and the second layer RMSA length is L_1 . Parametric study is done by keeping the dimensions constant of first layer of RMSA except G_1 which is 8mm and varying dimension length L_1 of second layer with three values 76 mm, 80 mm and 84 mm.

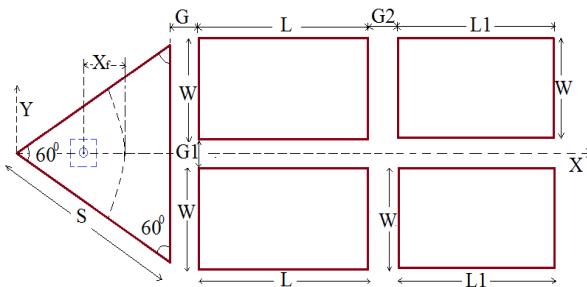


Fig.5. Proximity fed ETMSA gap coupled with four RMSA

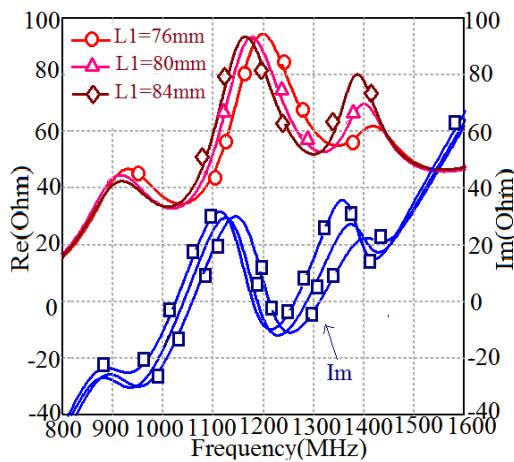


Fig.6. Resonance curve for ETMSA gap coupled with four RMSA

Fig.6. shows resonance curve for ETMSA gap coupled with four RMSA by varying length L_1 of RMSA of second adjacent layer. The effect of varying value of length L_1 on the coupling between RMSA of two layer is as shown in Fig.6.

TABLE II. PARAMETRIC STUDY OF ETMSA GAP COUPLED WITH FOUR RMSA

Sr.No	Length L_1 of second adjacent layer of RMSA in mm	Gap G_2 in mm	Feed X_f in mm	BW in MHz	Gain in dBi
1	76	8	22	626	9.5
2	80	8	22	632	9.53
3	84	8	22	629	9.51

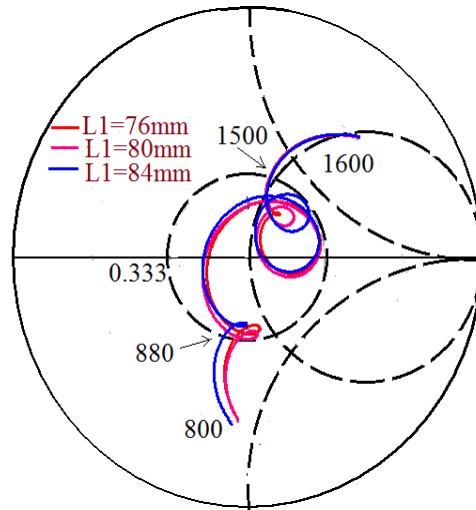


Fig.7. Impedance plot for ETMSA gap coupled with four RMSA

D. Proximity fed ETMSA gap coupled with five RMSA

This proposed configuration shown in Fig.8. is a three-layer structure with one ETMSA followed by two RMSA in the first layer and then three RMSA in the second adjacent layer. Gap between layers of RMSA is G_2 . Gap G_3 is 8mm.Gap G_3 is gap between the RMSAs in the second adjacent layer. Second adjacent layer RMSA width is W_1 and it is selected as 40mm to excite all three RMSA in the second adjacent layer.

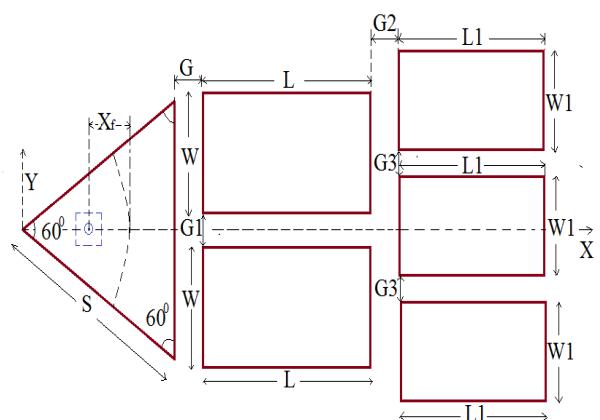


Fig.8. Proximity fed ETMSA gap coupled with five RMSA

Parametric study is done by keeping the dimensions constant of first layer of RMSA and varying dimension L2 of second layer with three values 76 mm, 80 mm and 84 mm. Optimum simulated BW obtained from the ETMSA gap coupled with five RMSA is 665 MHz (56%) and peak gain is 9.9 dBi for L1=90 mm and L2=80 mm. Fig. 9 shows impedance plot for ETMSA gap coupled with five RMSA.

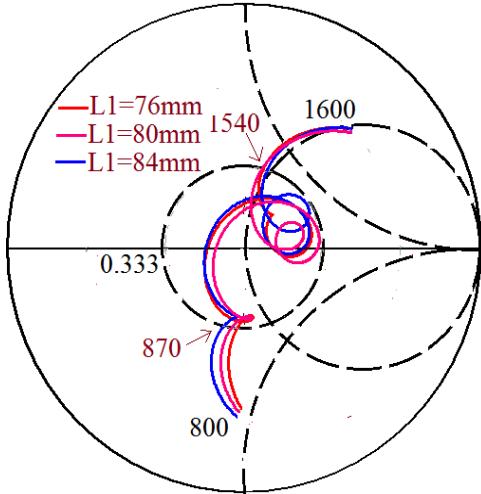


Fig.9. Impedance plot for ETMSA gap coupled with five RMSA

E. Proximity fed ETMSA gap coupled with six RMSA

This proposed configuration shown in Fig. 10 is a three-layer structure with one ETMSA followed by two RMSA in the first layer and then four RMSA in the second adjacent layer. Gap between two layer RMSA and four layer RMSA is G2 which is 8mm. Second layer structure consists of four RMSAs with the dimensions as inner two adjacent rectangles having length L2 and width W2. Two outer rectangles have length L1 and width W1. G3 is the gap between inner and outer rectangles patches.

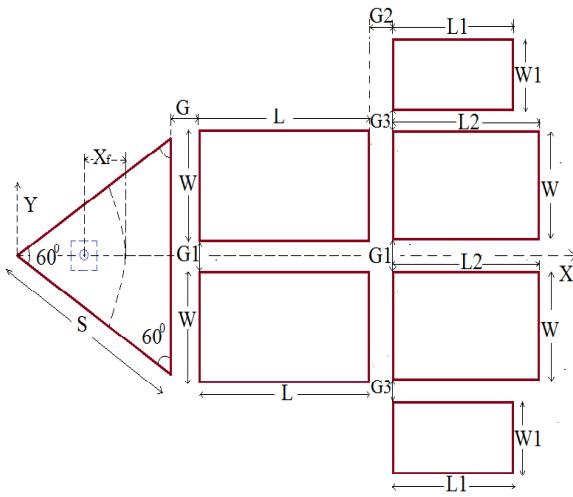


Fig.10. Proximity fed ETMSA gap coupled with six RMSA

Optimum simulated BW obtained from the ETMSA gap coupled with six RMSA is 672 MHz (57%) and peak gain is 9.7 dBi for L=90mm, L1=76mm, L2=80mm, W1=40mm. Fig.11 shows impedance plot for ETMSA gap coupled with six RMSA. The measured BW for ETMSA gap coupled with six RMSA is 680 MHz (57.5%).

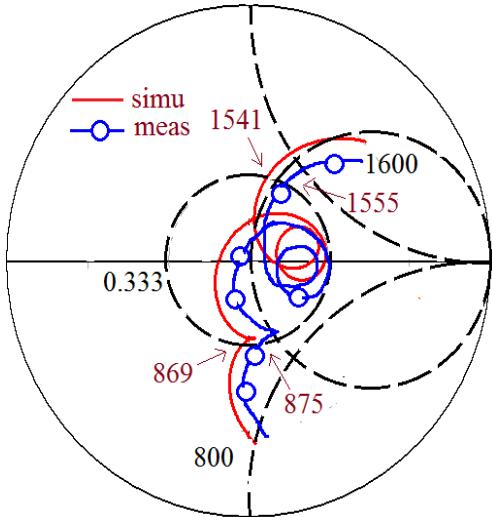


Fig.11. Impedance plot for ETMSA gap coupled with six RMSA

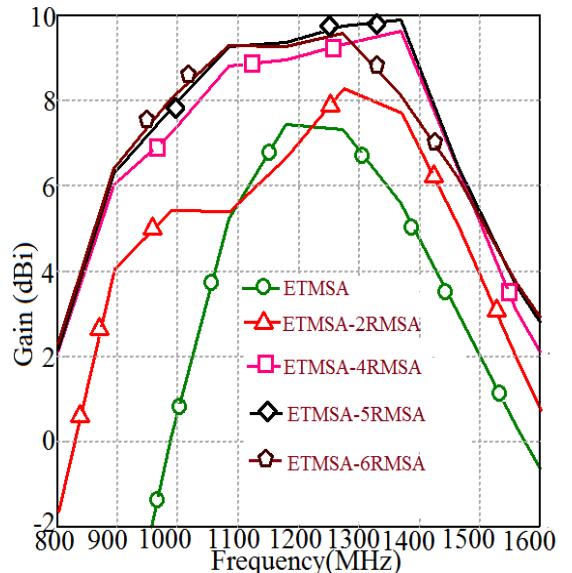


Fig.12.Simulated gain of proximity fed ETMSA gap coupled RMSAs

Fig.12 shows simulated gain of proximity fed ETMSA gap coupled with RMSAs. With variations of the structure using ETMSA gap coupled with RMSA, with increase in patches tapering, gain also increases. ETMSA gap coupled with six RMSA yields peak gain of 9.7 dBi due to illumination of more aperture area.

In these gap-coupled configurations, phase and geometric center of the configuration is shifted towards first gap-coupled rectangular patches. Hence maximum broadside gain is obtained in $\theta = 30^\circ$ direction. Gain remains maximum in the same direction throughout entire BW. Thus only modification in terms of antenna orientation is required from 0° to 30° in the θ direction. shows simulated and measured radiation pattern of ETMSA gap coupled with six RMSA at frequency 869 and 1540 MHz. All proposed configurations have peak gain of more than 7 dBi.

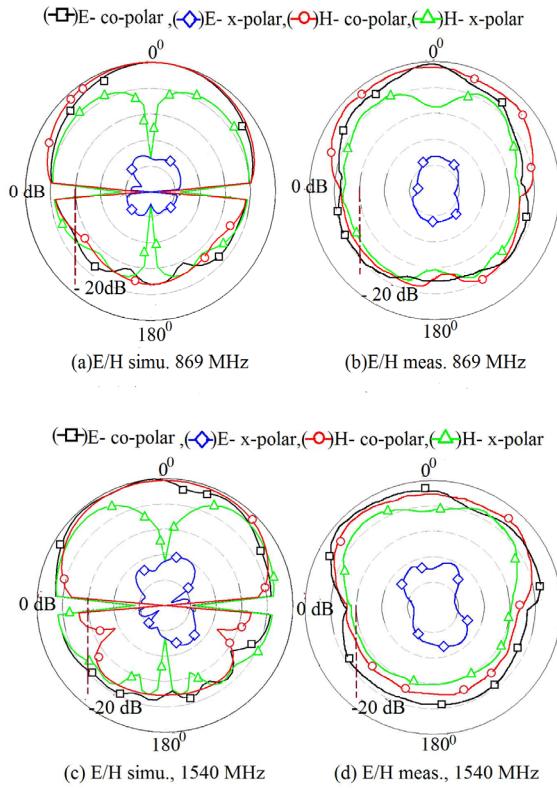


Fig.13. (a) Simulated (b) measured radiation pattern at 869 MHz, (c) simulated (d) measured radiation pattern at 1540 MHz of ETMSA gap coupled with six RMSA configuration

Fig.13 shows simulated and measured radiation pattern of ETMSA gap coupled with six RMSA configuration at frequency 869 MHz and 1540 MHz. High cross polar level in radiation pattern of ETMSA gap coupled with six RMSA configuration arises due to asymmetrical configuration and asymmetrical distribution of fields in both principal planes of ETMSA. Fig.14 shows fabricated prototype of proximity fed ETMSA gap coupled with six RMSA configuration. Ground plane of size 600mm x 600mm is used in the fabricated prototype.

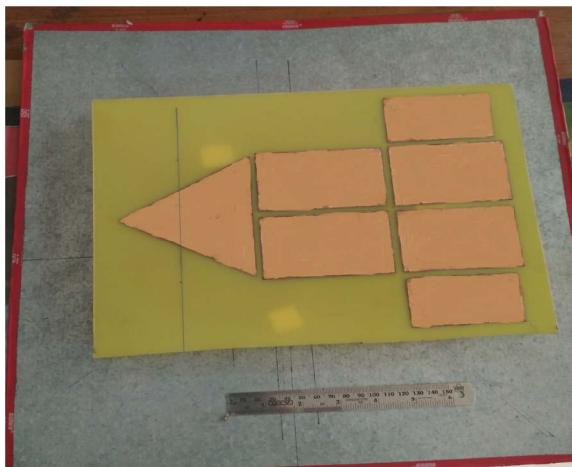


Fig. 14. Fabricated prototype of proximity fed ETMSA gap coupled with six RMSA

III. CONCLUSIONS

Proximity fed broadband gap coupled configurations of ETMSA with RMSA are proposed. The proposed configurations use multi-resonator technique to enhance the BW. The proposed configuration of ETMSA with double layer of RMSA with five RMSAs yields bandwidth of more than 665 MHz (56%) and peak gain of 9.9 dBi. Next proposed ETMSA gap coupled with six RMSAs yields bandwidth of more than 670 MHz (57%) and peak gain of 9.7 dBi at θ equal to 30° over the entire bandwidth.

REFERENCES

- [1] J. Abraham, T. Mathew and C. Aanandan, "A Novel Proximity Fed Gap Coupled Microstrip Patch Array for Wireless Applications," *Progress in Electromagnetics Research C*, vol. 61, pp. 171-178, 2016.
- [2] S. Hyuk Wi, Y. Shik Lee, and J. Gwan Yook, "Wideband Microstrip Patch Antenna With U-shaped Parasitic Element," *IEEE Transactions on Antennas and Propagation*, vol. 55, no. 4, 2007.
- [3] G. Kumar and K P. Ray, *Broadband Microstrip Antennas*, Norwood, MA Artech house, 2003.
- [4] A. Buffi, A. A. Serra, and P. Nepa, "A focused planar microstrip array for 2.4 GHz RFID readers," *IEEE Transactions on Antennas Propagation*, vol. 58, no. 5, pp. 1536-1544, May 2010.
- [5] K. H. Badr, "Design, fabrication, and measurement of four-by-four planar antenna sub-array," in Proc. 2nd Int. Conf. on Adaptive Science and Technology, 2009, pp. 396-401.
- [6] K.-S. Chin, H.T. Chang, J.-A. Liu H.-C. Chiu, J. S. Fu, and S.-H. Chao, "28-GHz patch antenna arrays with PCB and LTCC substrates" in Proc. Cross Strait Quad-Regional Radio Science and Wireless Technology Conf., Jul. 26-30, 2011, pp. 355-358
- [7] R. D Javor, X. D Wu, and K. Chang, "Design and performance of Microstrip Reflectarray Antenna", *IEEE Transactions on Antennas Propagation*, AP-43, no. 9, pp. 932-9, 1995.
- [8] R. Bhade and G. Kumar, "Dual Band Space-Fed Microstrip Antenna Arrays With Orthogonal Polarization," *Microwave and Optical Technology Letters*, vol. 53, No. 4, 2011.
- [9] R. Bhade and G. Kumar, "Equivalence of Space-fed Microstrip Antenna Array with Horn Antenna," *Microwave and Optical Technology Letters*, vol. 52, No. 5 May 2010.
- [10] A. Vaidya, R. Gupta, S. Mishra and J. Mukherjee "High-Gain Low Side Lobe Level Fabry Perot Cavity Antenna With Feed Patch Array," *Progress in Electromagnetics Research C*, vol. 28, pp 223-238, 2012.
- [11] G. DeJean, T. Thai, S. Nikolaou, and M. Tentzeris, "Design of Millimeter wave Microstrip Reflectarrays Design and Analysis of Microstrip Bi-Yagi and Quad-Yagi Array for WLAN Applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 6, 2007.
- [12] G. DeJean and M. Tentzeris, "A New High Gain Microstrip Yagi Array Antenna With a High Front to Back Ratio for WLAN and Millimeter wave applications" *IEEE Transactions on Antennas Propagation*, AP-55, No. 2, 2007.
- [13] O. Kramer, T. Djerafi, and W. Ke, "Very small footprint 60 GHz stacked Yagi antenna array," *IEEE Transactions on Antennas Propagation*, AP-59, no. 9, pp. 3204-3210, September 2011.
- [14] O. Kramer, T. Djerafi, and W. Ke, "Vertically Multilayer-Stacked Yagi Antenna with Single and Dual Polarizations," *IEEE Transactions on Antennas Propagation*, AP-58, no. 4, pp. 1022-1030, April 2010.
- [15] IE3D 12.1, Zeland Software, Freemont, USA.

Slot Cut Modified Triangular Shape Microstrip Antenna for Circular Polarization

Akshay V. Doshi

P.G.Student,
EXTC Department,
SVKM's DJSCOE,
Mumbai, India

doshiakshay4192@gmail.co

m

Amit A. Deshmukh

Prof. & Head,
EXTC Department,
SVKM's DJSCOE,
Mumbai, India

amit.deshmukh@djsce.ac.i

n

Sanjay B. Deshmukh

Research Scholar,
EXTC Department,
SVKM's DJSCOE,
Mumbai, India

sanjay.deshmukh@djsce.ac

.in

K.P.Ray

Prof. & Head
Electronics Department
DIAT,

Pune, India

kpray@rediffmail.com

Abstract— Modified design of triangular microstrip antenna obtained by combining two triangular shape patches is discussed. The offset distance between two patches tunes the spacing between orthogonal mode frequencies of the patch which yields circular polarized response with axial ratio bandwidth of 5%. Further to realize compact configuration, slot cut design of offset triangular overlap patches is studied. The rectangular slot helps in tuning the spacing between offset triangular patches along with offset distance to yield circular polarization. The detailed parametric study for the effects of slot along with offset distance in circular polarized antenna is presented. The tuning of orthogonal frequencies is obtained at lower offset distance between two patches for slot cut antenna. The optimum axial ratio bandwidth of 70 MHz in 1500 MHz frequency band is obtained along with broadside gain of 7 dBi and VSWR Bandwidth of 747 MHz. In optimum case slot cut overlapped design offers 42% reduction in patch area.

Keywords— Circularly polarized microstrip antenna; Broadband microstrip antenna; Modified Traingular microstrip antenna; Proximity feeding

I. INTRODUCTION

Circular polarized (CP) response is obtained when fields radiated by antenna have time and space orthogonality as well as they are equal in amplitude [1]. The orthogonal modes in MSA are generated by using nearly square patch in which TM_{10} and TM_{01} modes are closely spaced [2 – 4]. In rectangular geometries, circular MSA and equilateral triangular MSAs, CP response is obtained by cutting narrow slit in the patch [2 – 6]. In these MSAs single coaxial feed is used to excite orthogonal modes. Enhancement in axial ratio (AR) bandwidth (BW) as well as the gain is realized by using array of individual CP patches in which sequential rotation is used [7, 8]. Gain and BW have also been increased by using air suspended configurations [9, 10]. The CP response has also been realized by cutting widely used resonant slots like, U-slot, pair of rectangular slots, etc. [11 – 14]. E-shape MSA can also realize CP response [14]. In all these CP MSAs, feed point location is an important parameter since it decides magnitude of orthogonal current vectors of the excited modes. The circular polarized antennas have better gain (lesser than linear polarized antenna) and BW values as against shorted antennas. Therefore, in the communication applications wherein multipath propagation environment is present, CP antennas are preferred. The CP response is realized when two orthogonal resonant modes. The CP response has also been realized by cutting asymmetrical U-slot inside the square patch or by cutting symmetrical U-slot inside corner truncated patch

[13]. The CP response is also realized by cutting unequal length pair of rectangular slots on one of the edges of rectangular MSA (RMSA). In this paper, Slot cut circularly polarized modified triangular MSA (TMSA) in 1500 MHz frequency range is proposed. Slot is incorporated in modified triangular shape microstrip antenna [15] to realize CP response. For an offset proximity feed position, detailed parametric study for explaining the effects of varying slot length on varying offset distance between two patches and so on the excited resonant modes is presented. The geometry of slot incorporated and its center position is responsible to give CP response. Placement of proximity feed below the patch yields equal contribution for magnitudes of current vectors at orthogonal modes that further optimizes AR BW. For an offset distance of 2.2 cm between two TMSAs with slot of 0.4*4.6cm, VSWR BW of nearly 750 MHz (~50%) with AR BW of 70 MHz (25%) is obtained. This AR BW covers 750 MHz frequency range. Over a complete BW, radiation pattern is in the broadside direction. An optimum design yields broadside gain of more than 7 dBi over VSWR BW with peak gain of more than 8dBi.

Due to these antenna characteristics, proposed design can find application in mobile communication systems in 900 to 1500 MHz frequency band. The proposed antenna is first studied using IE3D simulations followed by experimental validation. The radiating patch is fabricated on glass epoxy substrate ($\epsilon_r = 4.3$, $h = 0.16$ cm, $\tan \delta = 0.02$) which is suspended above the ground plane using finite air gap ' Δ '. Antenna is fed using 50Ω N-type connector of 0.32 cm inner wire diameter. Square ground plane of 30 cm side length is used in simulations and measurements. In experiments, 'ZVH – 8' vector network analyzer, RF source 'FSC – 6' and spectrum analyzer 'SMB 100A' were used. 100A' were used. The measurements were carried out inside the Antenna Lab.

II. SLOT CUT MODIFIED TMSA

To achieve the compactness required to fulfil microstrip antenna characteristics and to obtain CP, in this paper we introduced the slot cut technique in reported work [15] which gives CP with higher AR BW and with lesser offset length. The parametric study is carried out for various offset lengths with various slots incorporated at centre of patch. Parametric study of offset length for modified TMSA is reported in [15], so in this paper we will see effect of slot on different offset lengths. At initial stage we will take bowtie configuration to study effect of slot first and then

similarly study is carried for different offset lengths. The steps involved in parametric study is given in following geometric configuration. (Fig.1).

We initially observe the resonance curve for bow-tie configuration. It is observed that TM_{10} & TM_{01} modes are closely spaced but it's not the optimum separation to realize optimum CP response [15]. Further looking it in detail it is clear that we need to reduce frequency of TM_{01} mode to get desired CP response hence slot should be placed along length so that effective width of current over patch surface increases which reduces the f_{01} . Hence by observing current distribution we have decided the slot length and position. With parametric variations we decide the slot length which yields maximum AR BW. Due to slot geometry TM_{01} mode gets disturbed as current along x-direction experiences larger path and hence f_{01} reduces. With horizontal slot introduced in star shape antenna it forces f_{10} frequency to reduce and come closer to f_{01} frequency. So, effect of horizontal slot is seen only in bow-tie configuration. Increase in f_{10} frequency is attributed to decrease in resonant length against the perturbation given by slots in the bow-tie MSA. The resonance curve plots, smith chart, current distribution & AR BW plots are given in Fig.2 (a-b) & Fig.3 (a-d). In further studies, above parametric variations are studied at different offset lengths. The geometric configuration is shown in Fig.4.

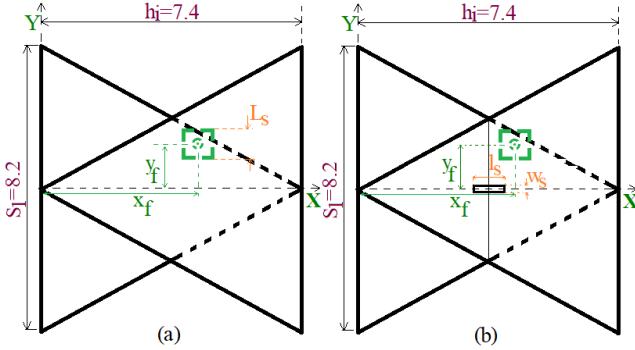


Fig.1: Geometrical configurations for parametric variations (a) Bow-tie without slot (b) Introduction of slot in Bow-tie

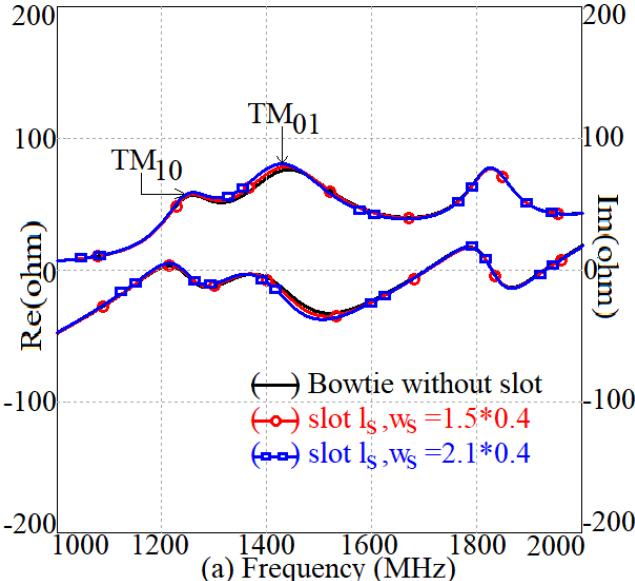


Fig. 2: (a) Resonance curve for bow-tie without slot & bow-tie slot cut microstrip antenna at 1500MHz

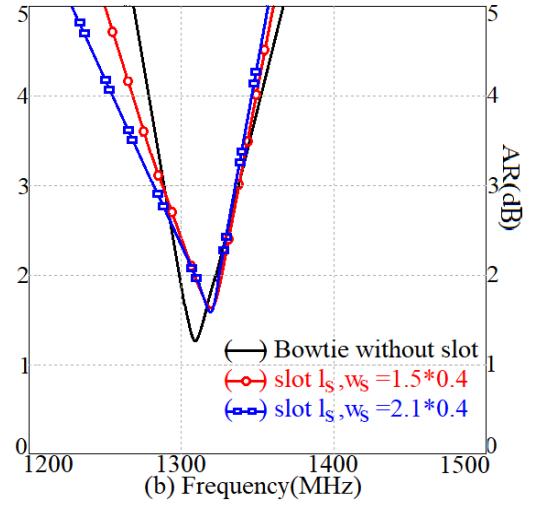


Fig. 2: (b) AR BW for bow-tie without slot & bow-tie slot cut microstrip antenna at 1500MHz

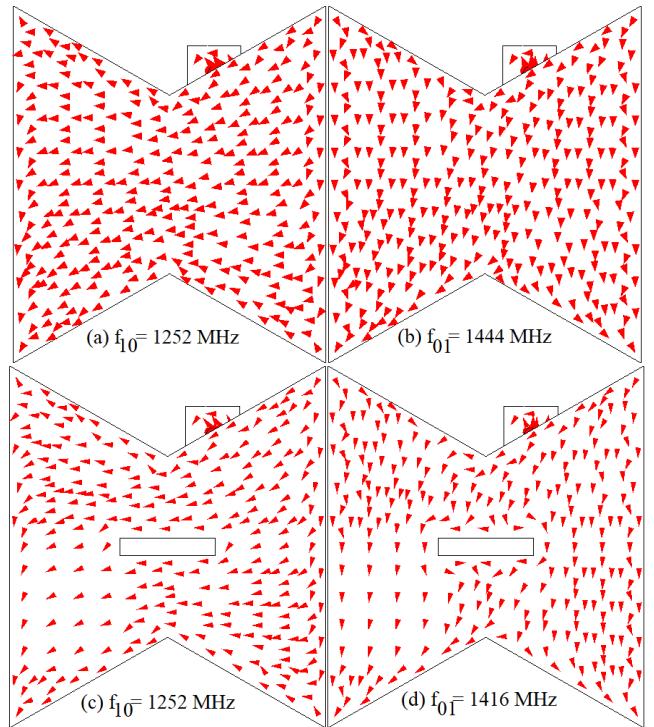


Fig. 3: Current distribution for (a, b): bow-tie without slot (c, d): bow-tie with slot

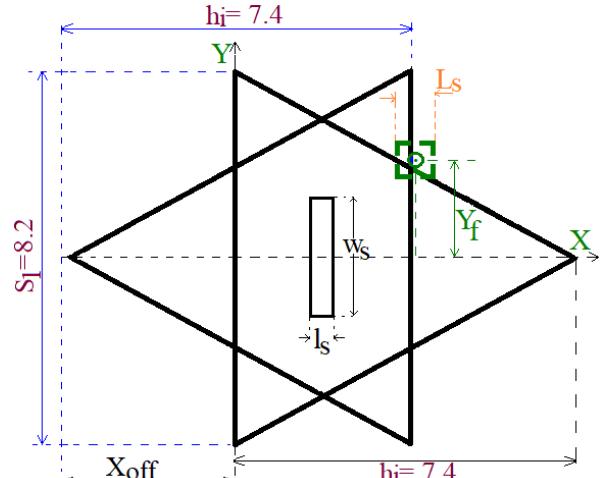


Fig.4: Configuration of star shape slot cut microstrip antenna at 1500MHz

A. Maintaining the Integrity of the Specifications

By observing resonance curve and current distributions we introduced vertical slots as vertical slot contributes to reduce effective length of TM_{01} mode it forces f_{01} to reduce and come closer to f_{10} hence realizing the CP response. Increase in f_{10} frequency is attributed to decrease in resonant length against the perturbation given by slots in the bow-tie MSA. Resonance curve and AR BW for optimum slot lengths at various offset lengths are given in Fig. 5 (a-b) & Fig. 6 (a-b).

The observations along results are given in tabular format. It is observed that required slot length to obtain CP is increasing up to 2.2cm offset length and then it reduces. This is because, as offset length increases two orthogonal modes start merging and after offset of 24mm they start separating out hence graph obtained is symmetrical. Plot of offset length verses slot length is given in Fig.7.

The parametric study is carried out and it is reported in tabular format. From Table1 it is clearly observed that we got CP response at specific offset length without slot. Hence reducing more than 40% effective aperture area. Optimum CP (max AR BW) obtained is also large compared to without slot configuration [15]. Resonance curve & AR BW is given in Fig. 8. The calculations for percentage area saving is given in Eq. (3) & Fig.10. Effective area for microstrip antenna is an area occupied by patch to radiate. Effective area calculation can be carried out with the help of Fig.10. The parametric study is carried out for offset length varying by 2mm. But reports of few offset lengths are presented to prove the concept. Radiation pattern is observed over entire VSWR BW. Pattern measurement is carried out in antenna lab. Simulated and measured radiation patterns are shown in Fig. 11 (a, b). The polarization plot is also given for final configuration to identify sense of polarization. As E_{left} $\phi=0, \phi=90$ having less than 3dB difference we can claim that this configuration offers LHCP.

In further section, the area occupied by modified TMSA without slot and the area covered by slot loaded modified TMSA is calculated to observe the compactness achieved by reported structure.

III. EFFECTIVE AREA CALCULATION

A. Effective area for modified TMSA to give promising AR BW:

From Fig.10 (a) effective area for modified TMSA can be calculated using following formulae:

$$\mathbf{A}_{e1} = \mathbf{A}_1 + 2\mathbf{A}_3 - 2\mathbf{A}_2 \dots \quad (1)$$

$$A_1 = \text{Area covered by rectangle} = 33 * 82 = 2706 \text{ sq.mm}$$

$$A_2 = \text{Area covered by triangular cut} = 150.81 \text{ sq.mm}$$

A_3 =Area covered by triangle due to offset length = 931.52 sq.mm

$$\text{So, } A_{el} = 4267.42 \text{ sq.mm}$$

This is the physical area covered by modified TMSA's radiating patch without introduction of slot. The effective area will be larger by $2\Delta l$ compared to physical

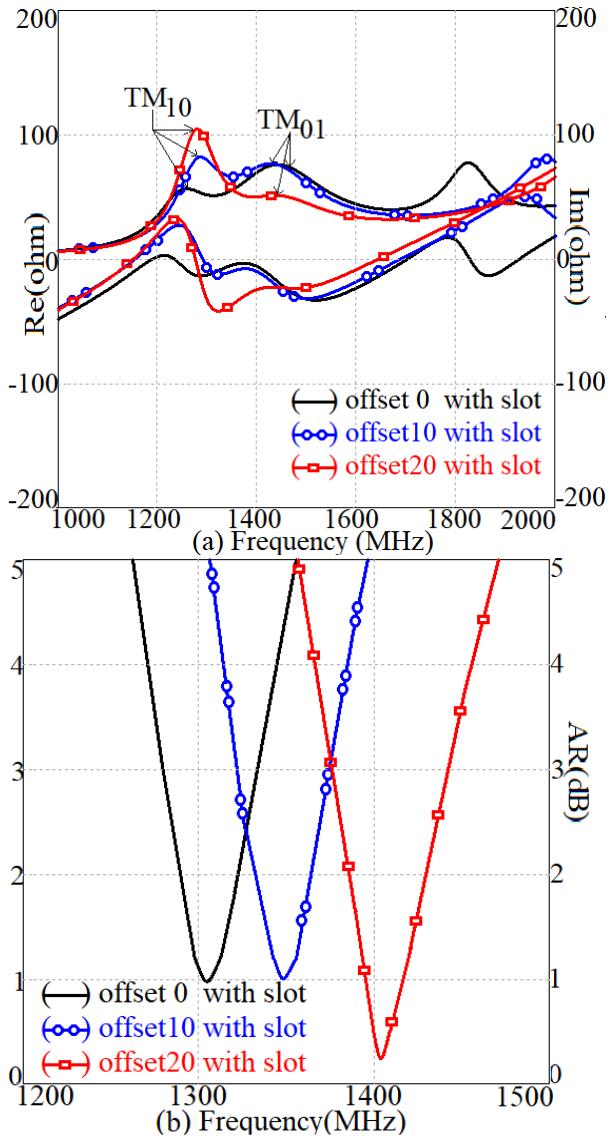


Fig. 5: (a) Resonance curves & (b) AR BW for slot cut star shape microstrip antenna for offset lengths 0, 10, 20

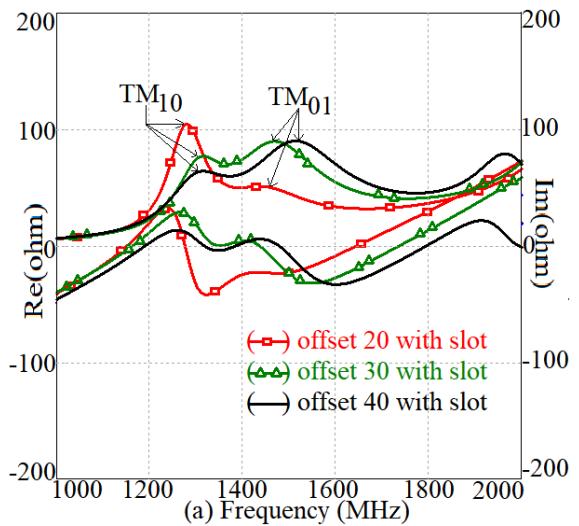


Fig. 6: (a) Resonance curves for slot cut star shape microstrip antenna for offset lengths 20, 30, 40

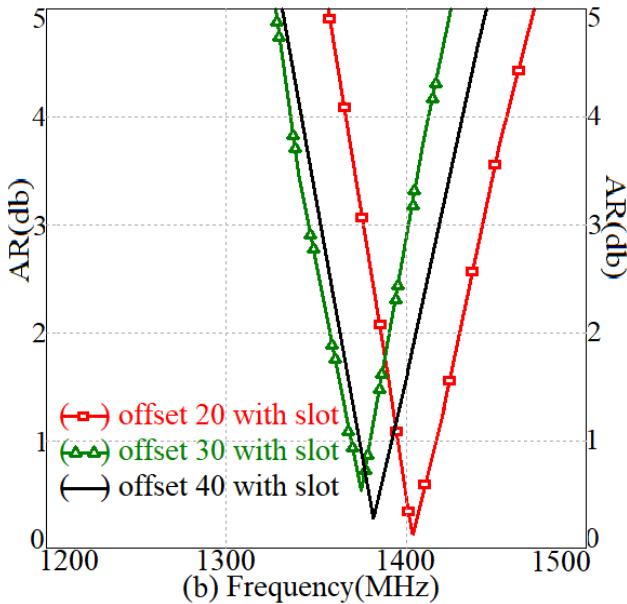


Fig. 6: (b) AR BW for slot cut star shape microstrip antenna for offset lengths 20, 30, 40

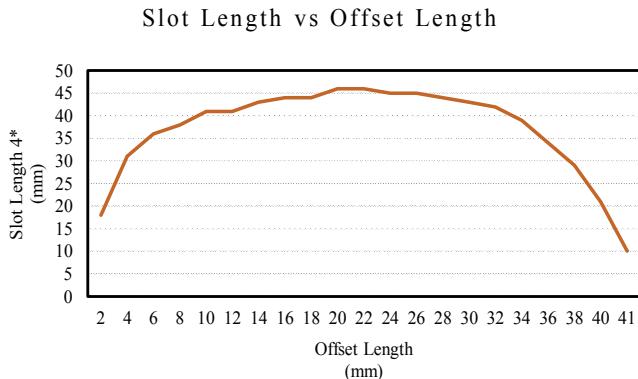


Fig. 7: Plot of offset length verses slot length

Table.1: Table of results for slot length at various offset length

Offset (cm)	VSWR BW without slot (MHz)	AR BW without slot (MHz)	Slot Length (cm)	VSWR BW with slot (MHz)	AR BW with Slot (MHz)
0	897.5	54.4	2.1*0.4	908	59.2
1.0	955	0	0.4*4.1	987	56
2.0	596	0	0.4*4.6	580	68.4
2.2	786	0	0.4*4.6	747	70.2
2.4	754	0	0.4*4.5	739	67.2
3.0	747	0	0.4*4.3	809	62.4
4.0	971	64.8	0.4*2.1	978	70
4.1	1012	67.2	0.4*1.0	1050	66

area, but to prove concept and for reducing complexity physical area is considered.

B. Effective Area for Slot cut Modified TMSA to give promising AR BW:

From Fig.10 (b) effective area for modified TMSA can be calculated using following formulae:

$$A_{e2} = A_1' + 2A_3' - 2A_2' - A_4' \dots\dots\dots(2)$$

$$A_{e2} = A_1' + 2A_3' - 2A_2' - A_4' \dots\dots\dots(2)$$

A_1 = Area covered by rectangle = $48*80 = 3840$ sq.mm
 A_2 = Area covered by triangular cut = 881.04 sq.mm
 A_3 = Area covered by triangle due to offset length = 277.64 sq.mm
 A_4 = Area of slot removed from patch = 184 sq.mm
 So, $A_{e2} = 2449$ sq.mm

C. Effective area for slot cut modified TMSA to give promising AR BW:

$$\% \text{Area Saving} = \frac{A_{e1} - A_{e2}}{A_{e1}} \times 100 \dots\dots\dots(3)$$

So percentage area saving is 42.61%. Fig. 12 (a, b) gives the comparison between modified TMSA without slot & modified TMSA with slot. It is clearly seen from AR BW plot that using slot cut techniques, required slot length to get CP response is less than modified TMSA without slot. It's also clear that the AR BW obtained is also more compared to modified TMSA. The simulated configuration is fabricated on glass epoxy substrate with thickness $h=1.6$ mm, dielectric constant $\epsilon_r=4.3$, $\tan \delta = 0.02$. Air suspended configuration with proximity feed technique is used for improvement of bandwidth with an air gap ' Δ ' = 1.8 cm. The total dielectric thickness is 1.96 cm. This configuration results in VSWR BW of 747 MHz (38.2%).

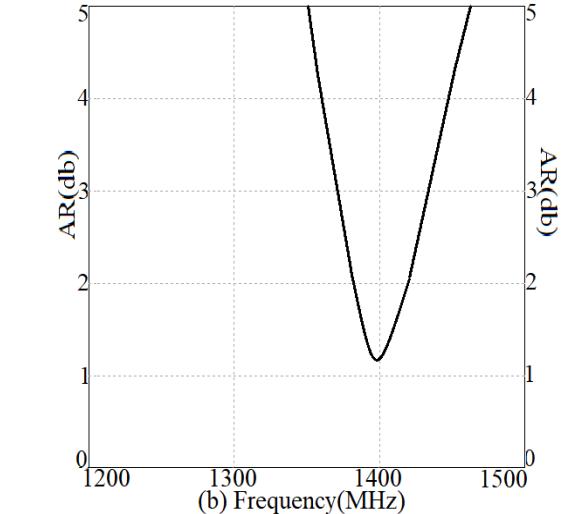
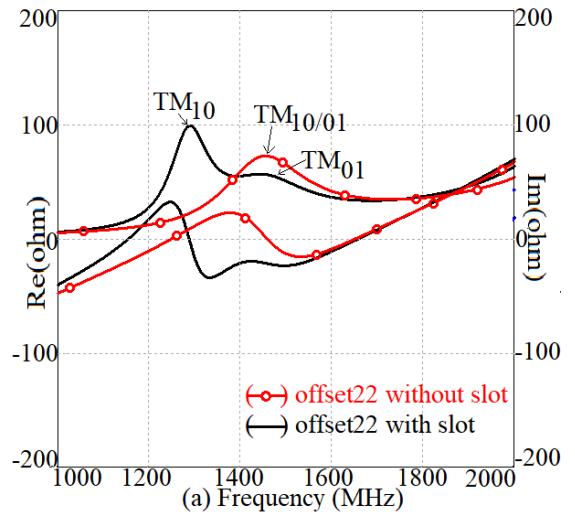


Fig. 8: (a) Resonance curve, (b) AR BW of modified TMSA for offset length of 22mm with & without slot.

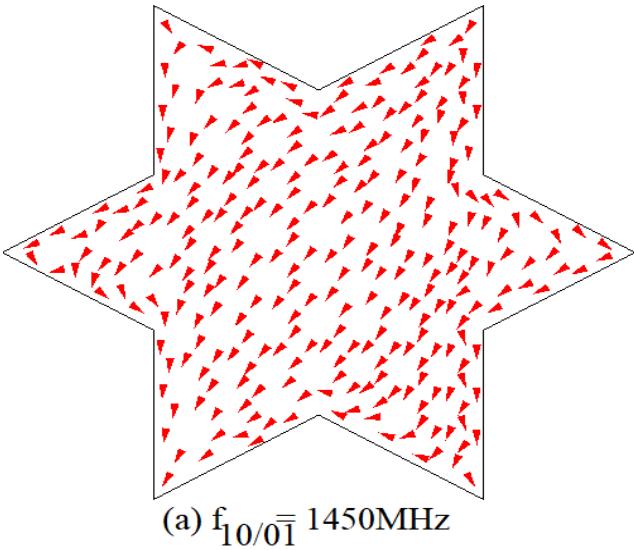


Fig. 9: (a) Current distribution of modified TMSA for offset length of 22mm without slot.

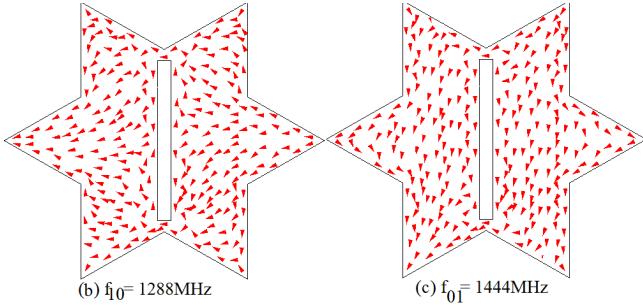


Fig. 9: (b-c) Current distribution of modified TMSA for offset length of 22mm with slot.

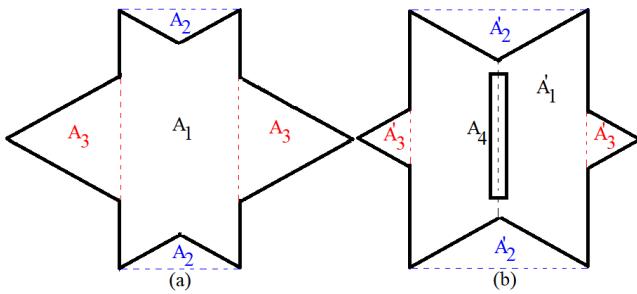


Fig. 10: Effective area for (a) Modified TMSA without slot (b) Slot cut modified TMSA

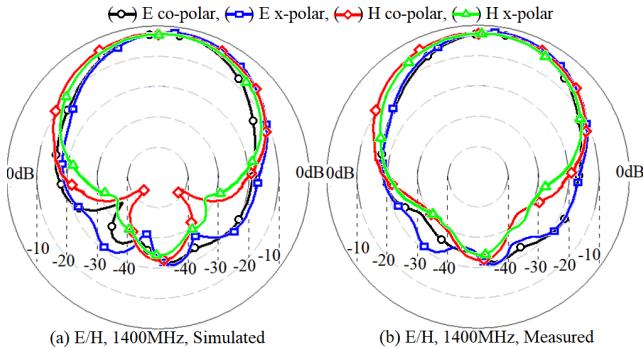


Fig. 11. (a-b) Radiation pattern for slot cut modified TMSA with offset distance 2.2cm.

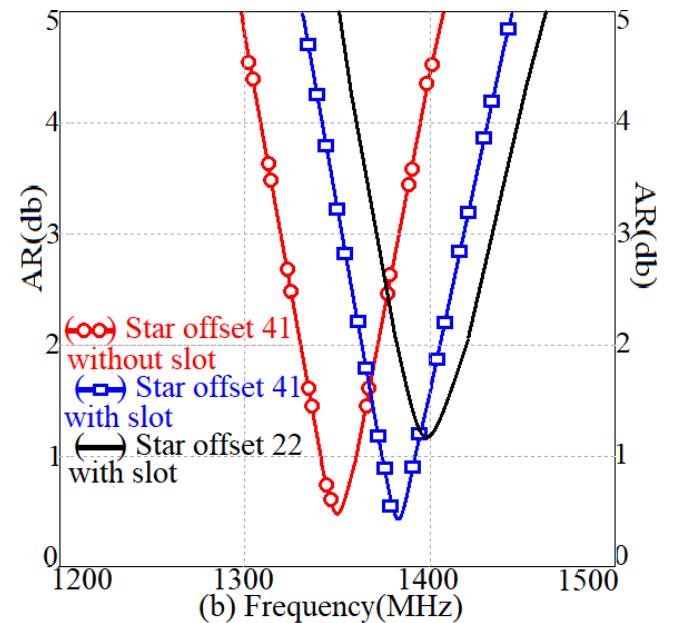
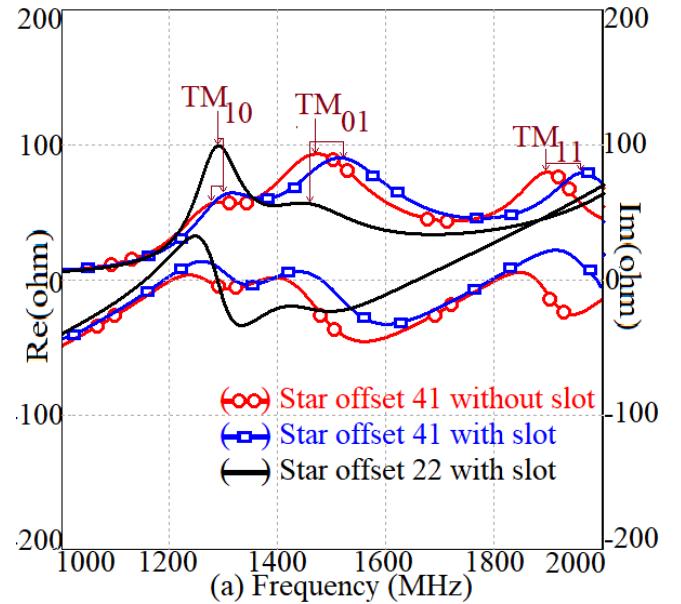


Fig. 12: (a) Resonance curve (b) AR BW for comparison between modified TMSA & slot cut modified TMSA.

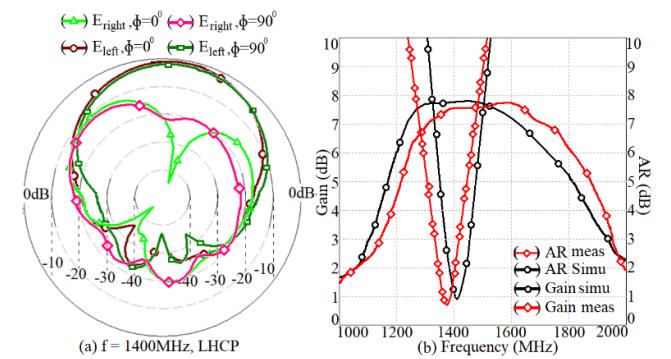


Fig. 13. (a) Polarization plot; (b) Gain and AR BW for slot cut modified TMSA with offset distance 2.2cm.

REFERENCES

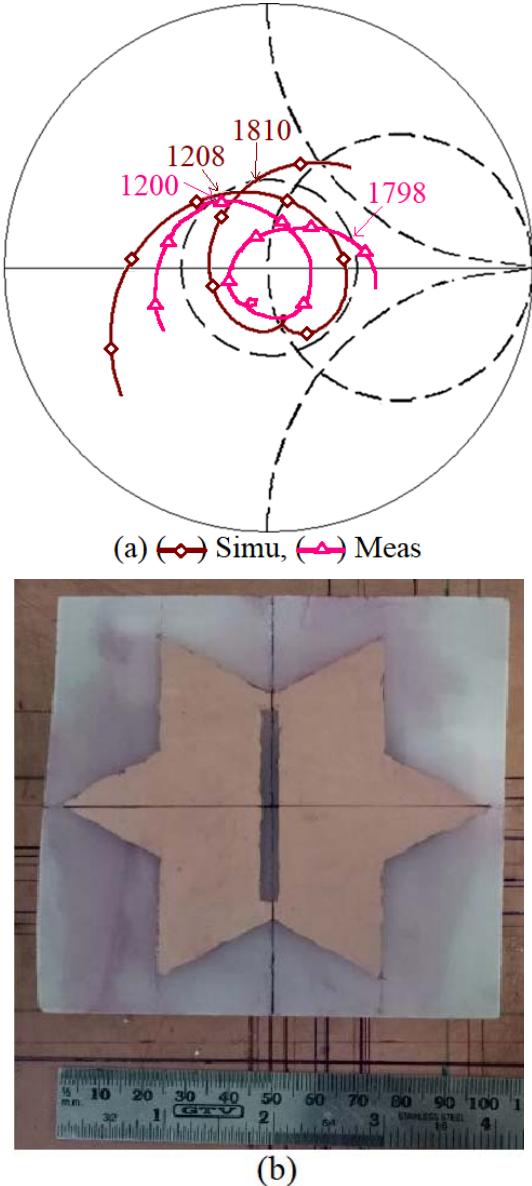


Fig. 14. (a) Impedance plot (b) Fabricated prototype of slot cut modified TMSA with offset distance 2.2cm

IV. CONCLUSIONS

Slot cut Modified TMSA will give better CP response than Modified TMSA by incorporating slot at center of patch. This will reduce radiating patch area which is termed as effective area. So by saving effective area compactness of patch is achieved. The path of current is increased by length of slot incorporated at center gives optimum spacing between orthogonal modes. A VSWR BW of more than 600 MHz ($>40\%$) with AR BW of 70 MHz (5%) is obtained. The proposed configuration shows broadside radiation pattern over complete BW with peak antenna gain of approximately 8 dBi. Proposed design can find applications in mobile communication systems in 900 to 1500 MHz frequency band.

- [1] Bhartia, B. and Bahl, I.J., *Microstrip antennas*, USA, 1980.
- [2] Garg, R., Bhartia, P., Bahl, I., and Ittipiboon, A *microstrip antenna design handbook*, Artech House, USA, 2001.
- [3] Kumar, G. and Ray, K.P., "Broadband microstrip antennas," Artech House, USA, 2003.
- [4] James, J. R., and Hall, P. S. "Handbook of microstrip antennas," Vol. 1, London: Peter Peregrinus, 1989.
- [5] Sharma, P. C., and Gupta, K. C., "Analysis and optimized design of single feed circularly polarized microstrip antennas," *IEEE Trans. Antennas Propagation*, 1983, AP-31, 6, pp. 949–955.
- [6] Deshpande, M. D., and Das, N. K., "Rectangular microstrip antenna for circular polarization," *IEEE Trans. Antennas Propagation*, 1986, AP-34, 6, pp. 744–746.
- [7] Wu, J., Yin, Y., Wang, Z., and Lian, R., "Broadband circularly polarized patch antenna with parasitic strips," *IEEE Antennas and Wireless Propagation Letters*, 2015, 14, pp. 559–562.
- [8] Morrow, I., and James, J. R., "Sequentially rotated large bandwidth circularly polarized printed antennas," *Electronics Letters*, 1995, 31, 24, pp. 2062–2064.
- [9] Deng, J., Guo, L., Fan, T., Wu, Z., Hu, Y., and Yang, J., "Wideband circularly polarized suspended patch antenna with indented edge and gap-coupled feed," *Progress In Electronics Research*, 2013, PIER 135, pp. 151–159.
- [10] Liu, N. W., Zhang, Z. Y., Zhao, J. Y., Fu, G., and Yao, Y., "A design of wideband circularly polarized antenna with high gain," *Microwave and Optical Technology Letters*, 2014, 56, pp. 1274–1277.
- [11] Khidre, A., Lee, K. F., Yang, F., and Eisherbeni, A., "Wideband circularly polarized e-shaped patch antenna for wireless applications," *IEEE Antennas and Propagation Magazine*, 2010, 52, 5, pp. 219 – 229.
- [12] Tong, K. F. and Wong, T. P., "Circularly polarized u-slot antenna," *IEEE Transactions on Antennas and Propagation*, 2007, 55, 8, pp. 2382 - 2385.
- [13] Lam, K. Y., Luk, K. M., Lee, K. F., Wong, H., and Bong, K., "Small circularly polarized u-slot wideband patch antenna," *IEEE Antennas and Wireless Propagation Letters*, 2011, 10.
- [14] Deshmukh Amit A., Zaveri Priyal, Deshmukh Sanjay, Odhekar Anuja, K.P.Ray, "Analysis of circularly polarized E-shaped microstrip antenna", APSYM, 2016.792155, DOI: 10.1109.
- [15] Deshmukh Amit A., Doshi Akshay, Kamble Pritish, and K. P. Ray, "Modified triangular shape microstrip antenna for circular polarization," 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India, Volume 115, 2017, pp. 101–107.
- [16] Deshmukh Amit A., Nagarbowdi Shafin, Kadam Poonam, Odhekar Anuja, "Broadband Gap-coupled Isosceles Triangular Microstrip Antennas," 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI) Pune Institute of Computer Technology, Pune, India, Feb 3-5, 2017.
- [17] Yang S. L. S., Lee K. F. and Kishk A. A., "Design and study of wideband single feed Circularly polarized microstrip antennas", *Progress In Electronics Research*, vol. 80, 2008, pp. 45–61.
- [18] Kumar, S., Kanaujia, B. K., Sharma, A., Khandelwal, M. K. and Gautam, A. K. (2014), "Single-Feed Cross-Slot Loaded Compact Circularly Polarized Microstrip Antenna For Indoor WLAN Applications", *Microw. Opt. Technol. Lett.*, 56: 1313–1317. doi: 10.1002/mop.28318
- [19] Hou, M.-J. and Row, J.-S. (2014), "Compact Circularly Polarized Microstrip Antenna For GPS Applications", *Microw. Opt. Technol. Lett.*, 56: 1293–1296. doi: 10.1002/mop.28335
- [20] S. Maddio, A. Cidronali, and G. Manes, "A NewDesign Method For Single-Feed Circular Polarization Microstrip Antenna With An Arbitrary Impedance Matching Condition," *IEEE Trans. Antennas Propag.*, vol. 59, no. 2, pp. 379-389,2011

A Circular Fractal Antenna Array

Rahul Chauhan

Dhirubhai Ambani Institute of
Information and Communication Technology
Gujarat, Gandhinagar-382007
Email: rahul871993@gmail.com

Sanjeev Gupta

Dhirubhai Ambani Institute of
Information and Communication Technology
Gujarat, Gandhinagar-382007
Email: sanjeev_gupta@daiict.ac.in

Abstract—In this paper, different arrays of the circular fractal antenna are analyzed. The various array design configurations are 1×2 , 1×4 , 2×2 and 2×4 . The substrate used is FR4 whose dimension is varied according to the array configurations. A 50Ω microstrip feed line is used to excite the antenna for all the configurations. The antenna shows stable radiation pattern and have good S_{11} characteristics. All the analysis is done in HFSS 2014.

Keywords- Fractal geometry, fractal antenna, arrays, microstrip feedline, multiband.

I. INTRODUCTION

THE evolution of mobile communication have seen the tremendous pace in the last 30 years and therefore the focus of researchers is shifted from traditional single band antenna to multiband antennas. Fractal seems to be the obvious choice because of its small size and excellent performance at wide range of frequencies. The term "fractal" was first coined by Mendelbort. Fractal geometries are preferred as they exhibit two properties a) Space-filling and b) Self-symmetry. Space-filling is used to miniaturize the antenna dimensions whereas self-symmetry is used to increase the bandwidth of the antenna. The fractal structure uses a virtual combination of capacitor and inductor due to which antenna resonates at different frequencies. As number of iterations increases in the fractal antenna, there is increase in the number of resonating frequencies and thus antenna shows multiband characteristics. Multi-band antennas have several applications in the wireless communication system such as cellular phones, Worldwide Interoperability for Microwave Access (WiMAX), Wireless Local Area Network (WLAN), radars, satellite communications, medical imaging, and surveillance.

In this paper, design and analysis of fractal antenna arrays is studied in detail and the performance of fractal antenna array is characterized in terms of S_{11} , gain and radiation pattern. The proposed antenna arrays have low fabrication cost, easy integration with other compact UWB (Ultra Wide Band) systems, small size and low profile.

II. FRACTAL ANTENNA AND ARRAY DESIGN

1) **Basic fractal antenna:** The basic fractal antenna is designed on FR4 (epoxy) substrate whose dimension is $48 \times 60mm^2$ and is iterated thrice.

Initially, the circular patch of radius 19mm is mounted on the substrate. To convert it into fractal shape 1 circle of radius 8mm in 1st iteration, 4 circles of radius 4mm in 2nd iteration, and 4 circles of radius 2mm in 3rd iterations are etched out from the substrate. The ground plane have dimension $48 \times 18mm^2$, and two squares of dimensions $5 \times 5mm^2$ and one rectangle having dimension $5 \times 4.42mm^2$ is etched out from the extreme ends and centre of the ground plane respectively. All this has been done in order to achieve UWB. The fractal antenna is operated at 3.4 GHz. The radius of the circular patch is calculated using the formula:

$$f_r = \frac{1.8412c}{2\pi a\sqrt{\epsilon_r}} \quad (1)$$

$$a = \frac{F}{\sqrt{1 + \frac{2h}{\pi\epsilon_r F} [\ln(\frac{\pi F}{2h}) + 1.7726]}} \quad (2)$$

$$F = \frac{8.791 \times 10^9}{f_r \sqrt{\epsilon_r}} \quad (3)$$

where

- c is the speed of light in free space
- f_r is resonating frequency
- a is radius of circular patch.
- h is height of substrate in cm.

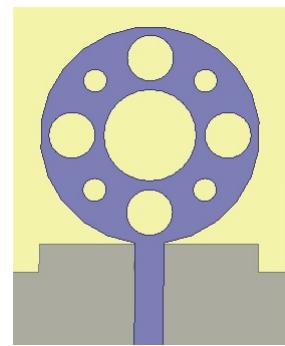


Fig. 1: Basic Structure of antenna

Figure 1 shows the basic structure of the antenna. Figure 2 and 3 shows the S_{11} and gain of the basic fractal antenna respectively.

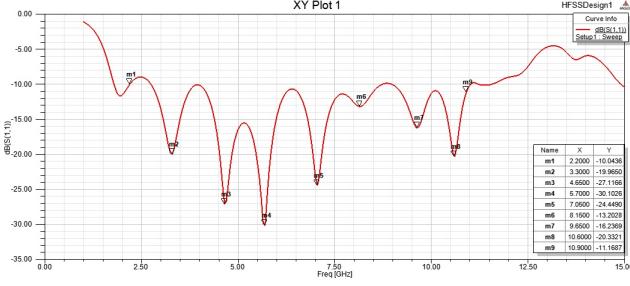


Fig. 2: S_{11} of basic fractal antenna

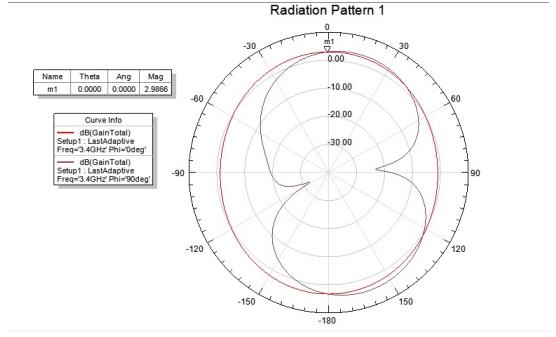


Fig. 3: Gain of basic fractal antenna

From the S_{11} plot (Figure 2), bandwidth is measured to be 8.7 GHz. The antenna resonates at frequencies 3.30 GHz, 4.65 GHz, 5.70 GHz, 7.05 GHz, 8.15 GHz, 9.65 GHz, 10.6 GHz, and have S_{11} values -9.96 dB, -27.11 dB, -30.10 dB, -24.44 dB, -13.20 dB, -16.23 dB and 20.33 dB respectively. The gain of the antenna is 2.98 dB and antenna shows stable radiation pattern. Radiation pattern is obtained by varying the ϕ and θ . All the array calculations have been done with the full ground plane because UWB is lost even if we etch out the ground plane.

2) **1 × 2 fractal array:** For two antenna elements to be independent, the inter-element spacing must be at-least $\lambda/2$. Since the operating frequency is 3.4 GHz the spacing comes out to be $44.11\text{mm} \approx 0.5\lambda$. Therefore placing the second element 44.11mm apart. These two fractal elements are connected with each other with $100\ \Omega$ line and a feedline of $50\ \Omega$ is used for the excitation of the antenna. However, there is a change in the dimensions of the substrate and ground plane. The dimension of the substrate is changed to $96 \times 60\text{mm}^2$ and the dimension of the ground plane is $96 \times 18\text{mm}^2$. Width of each feedline dependents on their impedance value and is calculated using the formula:

$$Z_c = \begin{cases} \frac{60}{\sqrt{\epsilon_{refl}}} \ln\left[\frac{8h}{w_0} + \frac{w_0}{4h}\right], & \frac{w_0}{h} \leq 1 \\ \frac{120\pi}{\sqrt{\epsilon_{refl}} \left[\frac{w_0}{h} + 1.393 + 0.667 \ln\left(\frac{w_0}{h} + 1.444\right) \right]}, & \frac{w_0}{h} > 1 \end{cases} \quad (4)$$

Figure 4 shows the structure of 1×2 fractal array.

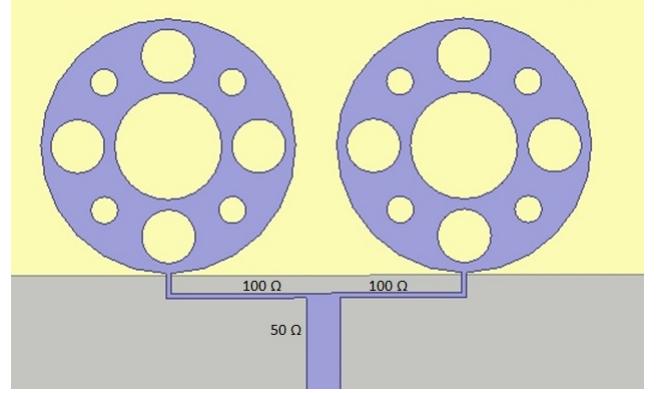


Fig. 4: 1×2 fractal array

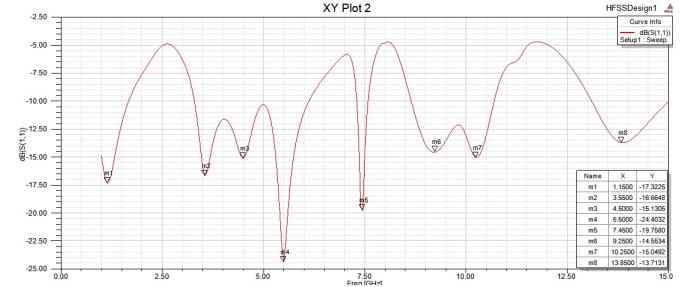


Fig. 5: S_{11} of 1×2 fractal antenna

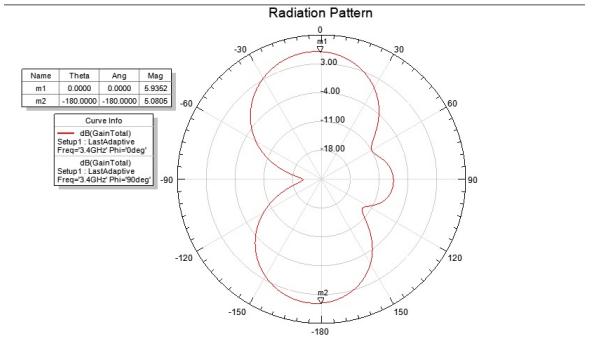


Fig. 6: Gain of 1×2 fractal antenna

Figure 5 and 6 shows the S_{11} and gain of the 1×2 antenna array. From figure 5, we can conclude that antenna is resonating at frequencies 1.15 GHz, 3.55 GHz, 4.50 GHz, 5.50 GHz, 7.45 GHz, 9.25 GHz, 10.25 GHz and 13.85 GHz and have S_{11} values -17.32 dB, -16.66 dB, -15.13 dB, -24.40 dB, -19.75 dB, -14.55 dB, -15.04 dB and -13.71 dB respectively.

The radiation pattern is obtained by varying ϕ and θ . The antenna has a gain of 5.93 dB. Therefore, we observe that

gain of the antenna increases significantly but UWB is lost and multibands are obtained. Thus, the fractal antenna array can be used as multi-band antenna.

3) **1×4 fractal array:** The inter-element spacing is 0.5λ , a quarter wave transformer is placed between a 50Ω equivalent point and 100Ω line whose resultant impedance is calculated using formula: $Z = \sqrt{Z_{in} * Z_{out}} = \sqrt{50 * 100} = 70.7\Omega$. Also, 70.7Ω is further connected to 100Ω line whose equivalent is combined with 50Ω microstrip feedline. The dimension of the substrate is changed to $192 \times 60mm^2$ and ground plane dimension is changed to $192 \times 18mm^2$. Figure 7 shows 1×4 fractal array structure.

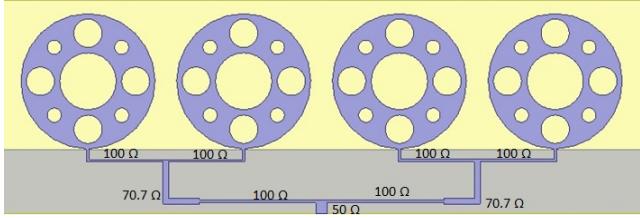


Fig. 7: 1×4 fractal antenna array

Figure 8 and 9 shows S_{11} and gain of the 1×4 fractal array. From figure 8 we can conclude that antenna is resonating at frequencies 4.70 GHz, 6.15 GHz, 8.30 GHz, 9.70 GHz, 10.95 GHz, 12.75 GHz and 14.25 GHz and have S_{11} values -31.67 dB, -24.51 dB, -14.56 dB, -27.04 dB, -45.12 dB, -25.24 dB, and -13.65 dB respectively. The antenna has a gain of 8.08 dB. Therefore it can be observed that gain of 1×4 antenna configuration further increases significantly but, at the cost of increase in side lobes.

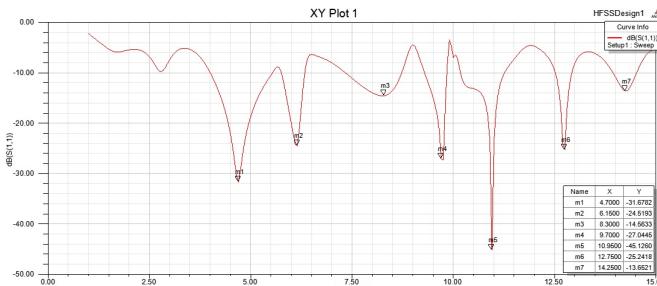


Fig. 8: S_{11} of 1×4 array

4) **2×2 fractal array:** A 2×2 fractal array which is also called as a corporate array is made by keeping inter-element spacing of 0.5λ in the adjacent element and by increasing dimensions of the substrate to twice that of the basic fractal substrate. Also, the 100Ω line is used to connect two adjacent elements, and feedline of 50Ω is attached to excite the antenna. The dimension of the substrate is $96 \times 120mm^2$

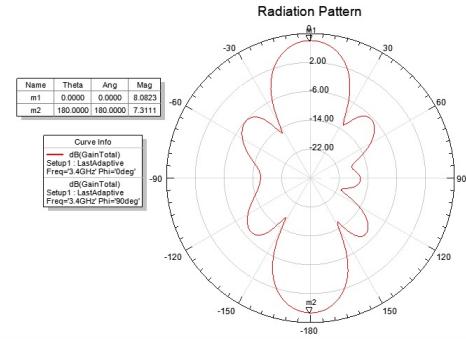


Fig. 9: Gain of 1×4 array

whereas ground plane has the dimension $96 \times 18mm^2$. Figure 10 shows the structure of 2×2 fractal antenna array.

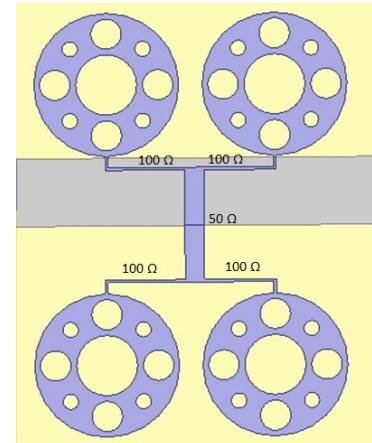


Fig. 10: 2×2 fractal antenna array

Figure 11 and 12 shows the S_{11} and radiation pattern of the given structure. From figure 11, we can conclude that antenna is resonating at frequencies 3.55 GHz, 4.45 GHz, 5.40 GHz, 7.40 GHz, 8.95 GHz, 9.60 GHz and 13.40 GHz with the S_{11} values -14.89 dB, -15.33 dB, -38.12 dB, -20.08 dB, -14.21 dB, -14.77 dB, and -16.10 dB respectively. The antenna has a gain of 5.94 dB.

Therefore, if we compare 4 element array i.e. linear (1×4) array and corporate(2×2) array we observe that the linear array have significant higher gain as compared to the 2×2 corporate array but with a drawback that there are more side lobes in the 1×4 array as compared to the 2×2 array.

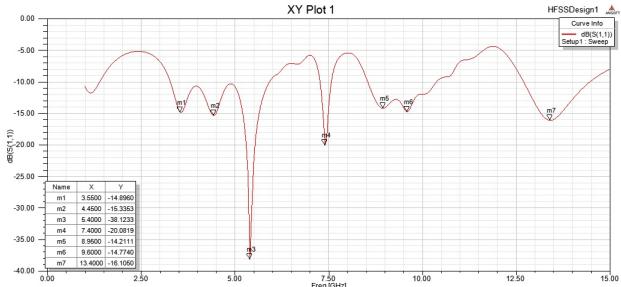


Fig. 11: S_{11} of 2×2 array

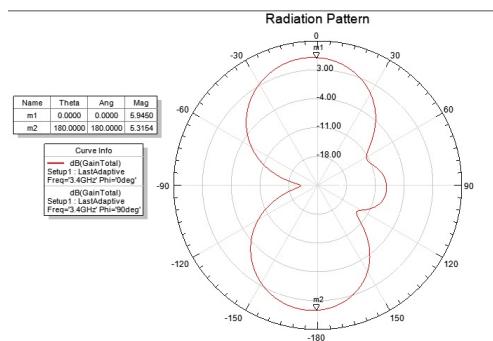


Fig. 12: Gain of 2×2 fractal antenna

5) **2×4 fractal array:** Similar to the 1×4 linear array, a 2×4 corporate array is made using quarter wave transformer. The dimension of the substrate and ground plane is $96 \times 240\text{mm}^2$ and $96 \times 18\text{mm}^2$ respectively. Figure 13 shows the structure of the 2×4 corporate array.

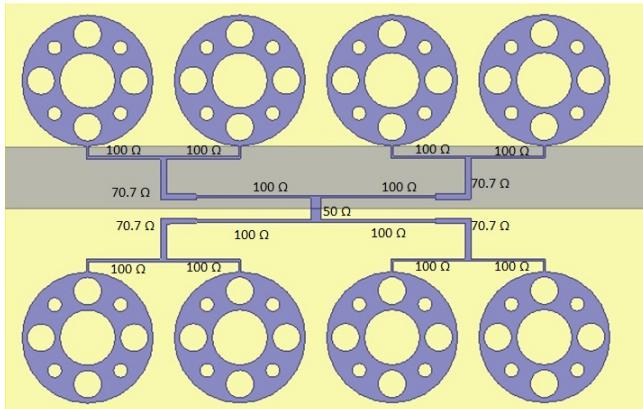


Fig. 13: 2×4 fractal antenna array

The S_{11} and radiation pattern are given in figure 14 and 15 respectively. From figure 14, we can observe that antenna resonates at frequencies 2.75 GHz, 3.80 GHz, 4.45 GHz, 5.20 GHz, 6.00 GHz, 6.90 GHz, 7.70 GHz, 8.40 GHz, 9.40 GHz,

10.75 GHz, 12.85 GHz, 13.65 GHz and 14.25 GHz and have S_{11} values -25.02 dB, -18.06 dB, -16.50 dB, -11.94 dB, -20.66 dB, -11.04 dB, -14.83 dB, -10.71 dB, -14.76 dB, -22.22 dB, -20.89 dB, -37.56 dB and -24.85 dB respectively.

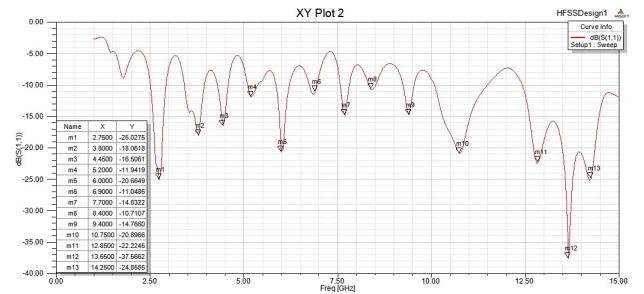


Fig. 14: S_{11} of 2×4 array

The gain of the antenna is 8.53 dB (Figure 15). Thus gain of 2×4 array is slightly increased as compared to 1×4 array, although resonating frequencies are more in 2×4 array than in 1×4 array.

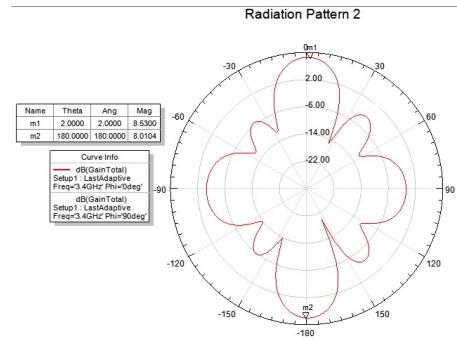


Fig. 15: Gain of 2×4 array

III. FABRICATION AND TESTING

Two fractal design, basic fractal antenna and 1×4 fractal antenna array were practically implemented and results were measured with the help of Vector Network Analyzer (E5071B ENA Series) in the frequency range of 1-8 GHz.

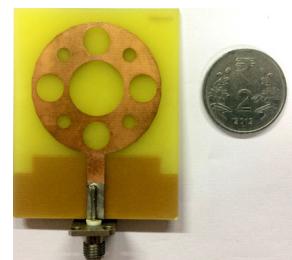


Fig. 16: Basic fractal antenna



Fig. 17: Bottom view of basic fractal antenna

Figure 16 and 17 shows the top and bottom view of the basic fractal antenna respectively. In figure 18, device can be seen under test. Figure 19 shows the practical S_{11} values of basic fractal antenna. The antenna resonates at the frequencies 2.04 GHz, 5.33 GHz, 7.44 GHz and have S_{11} values -20.76 dB, -37.54 dB and -31.44 dB respectively (simulated values are 3.30 GHz, 5.70 GHz, and 7.05 GHz).



Fig. 18: Testing of basic fractal antenna using VNA

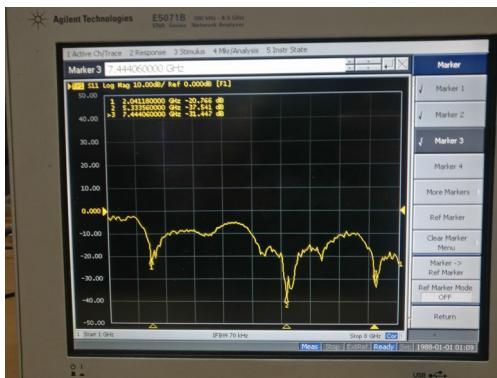


Fig. 19: S_{11} for basic fractal antenna

Also, 1×4 fractal antenna array is practically realized. Figure 20 and 21 shows the top and bottom view of 1×4 fractal antenna respectively. In figure 22 device can be seen under test. From figure 23, it is observed that antenna is resonating at the frequencies 4.53 GHz, 6.06 GHz with S_{11} values -15.59 dB and -32.65 dB respectively (simulated values are 4.70 GHz and 6.15 GHz)

Thus we observe that practical and simulated values are almost identical for the frequency range of 1-8 GHz with very slight deviation due to coaxial feed which is required in the practical design for the excitation of the antenna.



Fig. 20: Top view of 1×4 fractal antenna array

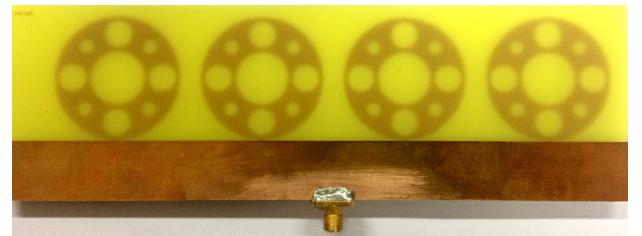


Fig. 21: Bottom view of 1×4 fractal antenna array



Fig. 22: Testing of 1×4 fractal antenna array using VNA

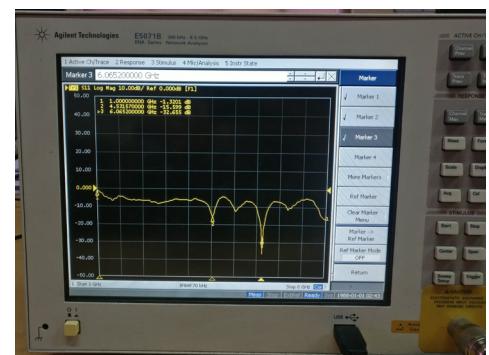


Fig. 23: S_{11} for 1×4 fractal antenna array

IV. CONCLUSION

In this paper, various configurations of the fractal antenna array were discussed. Also, it was observed that linear configurations of the fractal array (1×2 and 1×4) have more gain as compared to the corporate array configurations (2×2 and 2×4), but with the drawback that the side lobe level increases in the linear array configurations. Table 1 shows the gain comparison for different array configurations.

Array Configuration	Gain(dB)
Basic Fractal	2.98
1×2 array	5.93
1×4 array	8.08
2×2 array	5.94
2×4 array	8.53

TABLE I: Gain of different array configurations

As we move from basic fractal antenna to array configuration we observe that UWB is lost, but there is significant increase in the gain and antenna shows multiband characteristic. Thus an array of fractal antenna can be used as multi-band antenna, as it covers wide range of frequencies.

REFERENCES

- [1] I. Acharya and D. Upadhyay, "A Novel Circular Fractal Antenna with Band Notch Characteristics for UWB Applicatons," 2015 Fifth International Conference on Advances in Computing and Communications (ICACC), Kochi, 2015
- [2] D. H. Werner and S. Ganguly, "An overview of fractal antenna engineering research," in IEEE Antennas and Propagation Magazine, vol. 45, no. 1, pp. 38-57, Feb. 2003.
- [3] A. A. Potapov, "The base of fractal antenna theory and applications: Utilizing in electronic devices," 2013 IX International Conference on Antenna Theory and Techniques, Odessa, 2013
- [4] Savina Neetu Bansal R K Bansal "Design and Analysis of Fractal Antennas based on Koch and Sierpinski Fractal Geometries" International Journal of Advanced Research in Electrical Electronics and Instrumentation Engineering vol. 2 no. 6 June 2013.
- [5] E. L. Barreto, A. G. d'Assuno and L. M. Mendona, "A new fractal antenna array for wireless communications," 2015 31st International Review of Progress in Applied Computational Electromagnetics (ACES), Williamsburg, VA, 2015.
- [6] Zhen Yu, Jianguo Yu, Xiaoying Ran, and Chenhua Zhu, A Novel Ancient Coin-Like Fractal Multiband Antenna for Wireless Applications, International Journal of Antennas and Propagation, vol. 2017, Article ID 6459286, 2017.
- [7] Madi, MA, Al-Husseini, M, Ramadan, A, Kabalan, KY and El-Hajj, A. 2012. A reconfigurable cedar-shaped microstrip antenna for wireless applications. Progress in Electromagnetics Research C, 25: 209221.
- [8] D. Upadhyay, R. P. Dwivedi, "Antenna miniaturization techniques for wireless applications", Eleventh IEEE International Conference on Wirelessand Optical Communication Networks (WOCN), 2014.
- [9] X. Yang J. Chiochetti D. Papadopoulos L. Susman "Fractal antenna elements and arrays" Appl. Microw. Wireless vol. 5 no. 11 pp. 34-46 May 1999.
- [10] D. H. Werner, R. L. Haupt and P. L. Werner, "Fractal antenna engineering: the theory and design of fractal antenna arrays," in IEEE Antennas and Propagation Magazine, vol. 41, no. 5, pp. 37-58, Oct. 1999.
- [11] Kumar Raj, Magar Dhananjay, K. KailasSawant, "On the design of inscribed circular fractal antenna for UWB applications", International Journal of Electronics and Communications (AEU), vol. 66, pp. 68-75, 2012.
- [12] Antenna Theory: Analysis and Design 3rd edition, by Constantine A. Balanis.

High Performance Multiplierless Serial Pipelined VLSI Architecture for Real-Valued FFT

Jinti Hazarika¹, Mohd Tasleem Khan², Shaik Rafi Ahamed³, and Harshal B. Nemade⁴

Department of Electronics and Electrical Engineering

Indian Institute of Technology Guwahati

jinti@iitg.ac.in¹, tasleem@iitg.ac.in², rafiahamed@iitg.ac.in³, harshal@iitg.ac.in⁴

Abstract—This paper presents a high-performance multiplierless serial pipelined architecture for real-valued fast Fourier transform (FFT). A new data mapping scheme (DMS) is suggested for the proposed serial pipelined FFT architecture. The performance is enhanced by performing FFT computations in $\log_2 N - 1$ stages followed by a select-store-feedback (SSF) stage, where N is the number of points in FFT. Further enhancement in performance is achieved by employing quarter-complex multiplierless unit made up of memory and combinational logic in every stage. The memory stores half number of partial products while the remaining partial products are taken care by external combinational logic. Compared with the best existing scheme, the proposed design reduces the computational workload on half-butterfly (H-BF) units by $(2N - 8)$. Application specific integrated circuit (ASIC) and field programmable gate array (FPGA) results show that the proposed design for 1024-point achieves 31.54% less area, 30.13% less power, 33.56% less area-delay product (ADP), 27.11% less sliced look-up tables (SLUTs) and 28.37% less flip-flops (FFs) as compared to the best existing scheme.

Index Terms—Fast Fourier transform (FFT), real-valued signals, multiplierless, pipelined architecture, serial commutator.

I. INTRODUCTION

FAST Fourier transform (FFT) plays a vital role in a varied number of applications such as wireless communications, image and signal processing applications. It is a fundamental transform technique in digital signal processing (DSP) which is being extensively researched at both algorithmic and architectural levels. Over the years, researchers have made continuous efforts for improving the performance of FFT processors in terms of throughput, area and power consumption. In particular, low-area and low-power FFT architectures are in demand since most of the portable devices have a limited power supply, and are expected to consume as less area as possible.

On the basis of intermediate data processing, FFT architectures are mainly classified into two main categories: memory-based and pipelined. In case of memory-based architectures, latency due to stage-wise computations with the same processing element is significant, and dedicated memory addressing scheme is required to make it continuous-flow. To overcome these issues, several approaches have been reported [1], [2]. For instance, a new stage partition scheme was proposed in [1] which shifts the twiddle factors

and the butterflies to be within one stage to reduce the number of computation cycles, another scheme [2] based on the partition scheme in [1] was suggested to obtain a normal order output and continuous-flow. In the recent past, pipelined architecture is gaining attention towards efficient implementation of FFT. More specifically, pipelined architecture gets the priority when high-throughput and low-power are desired due to its ability to process continuously. On the basis of samples processed per iteration, these can be divided into two categories: serial (or single-path) and parallel (or multi-path) architectures. Serial FFT algorithms [2], [3] are preferred for low-power applications since it offers lower usage of computational resources over its parallel counterparts. However, the presence of multiplier unit is still major concern from power consumption point of view. It is a known fact that the number of multiplications required for FFT computation varies logarithmically with the number of points. In practice, large point FFT is always desired which would result in higher power consumption, and thereby making its real-time implementation difficult.

In the past, a number of FFT designs have been presented in literature to reduce the chip-area consumed by multipliers [4]–[6], for instance, co-ordinate rotation digital computer (CORDIC)-based designs [4], distributed arithmetic (DA) based designs [6], etc. However, there is always an accuracy problem with designs based on co-ordinate rotation digital computer (CORDIC) since it takes several iterations to reach the desired result. On the other hand, DA based design is primarily suitable for computing larger length inner products. High radix implementation of the aforementioned techniques could be used for minimizing the multiplications in FFTs [7], but increases the system complexity. Different approaches have also been presented to eliminate the complex multiplier completely by special constant coefficients using shifters and adders. This results in significant reduction in number of operations, hence the power consumption reduces but this introduces delay into the system. To overcome the above mentioned issues, memory based implementation of multiplier would be an alternative approach. With advancements in complementary metal oxide semiconductor (CMOS) technology scaling, the cost of memory has reduced making it faster and more power-efficient over its predecessors.

According to international technology roadmap for semi-

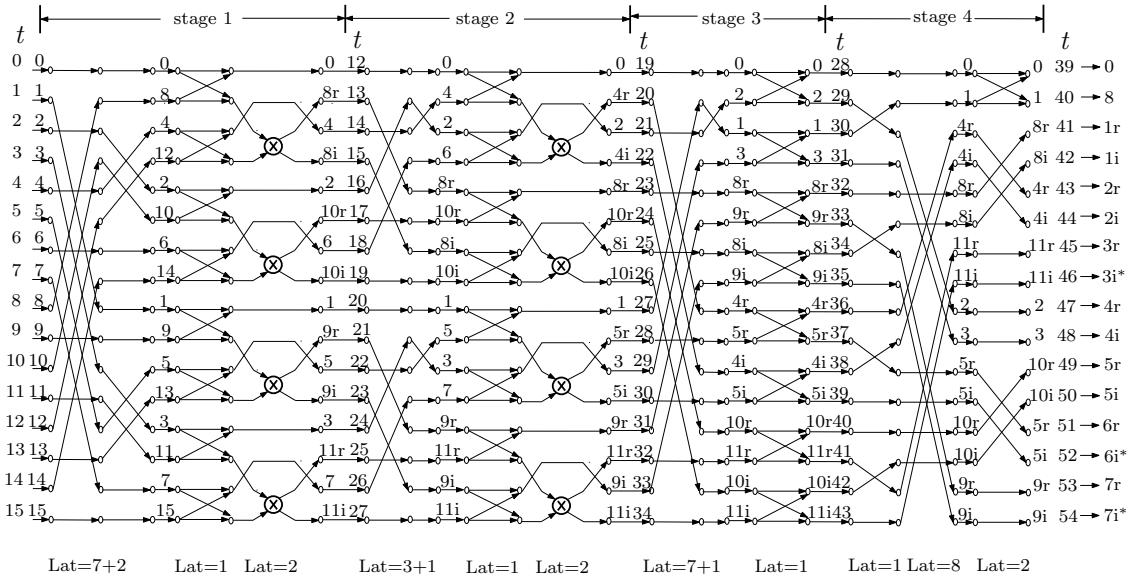


Fig. 1. Proposed data mapping scheme (DMS) for a 16-point RFFT.

conductors (ITRS), memory will have dominating presence in the future system-on-chip (SoC). The reason of this explanation is transistor density in logic cells is found to be lower than the SRAM cells [8]. It also provides high-throughput and reduced-latency implementation due to shorter memory-access-time than usual multiplication-time. Nonetheless, it is expected that memory would consume less dynamic power consumption due to less switching activities for memory-read operations over conventional multiplier structure. The key contributions of this work are listed as follows:

- Data mapping scheme (DMS) for the serial-pipelined FFT is proposed.
- Novel select-store-feedback (SSF) unit for the last stage of FFT computation is proposed.
- New design for memory based multiplier (MBM) with external combinational logic is provided.

The rest of this paper is organized as follows: Section II discusses about the proposed real-valued FFT. Section III presents the architectural details of proposed serial-pipelined FFT. Section IV compares the proposed and existing designs in terms of hardware complexity, area and power. Section V gives the conclusion of the presented work.

II. PROPOSED REAL-VALUED FFT

In case of pipelined FFT architectures, dedicated DMS is required for processing of input and intermediate data through the FFT processor. In every stage, data re-ordering (DR) is required for mapping the intermediate data, according to the data flow. In order to increase the efficiency of serial-pipelined FFT computation, it is necessary to enhance the resource utilization of butterfly (BF) units and reduce the multiplier complexity. A new DMS is suggested for serial-pipelined FFT for enhancing the resource utilization of BF units while the MBM is suggested for reducing the complexity of twiddle factor multiplication. It is important

to note that DR circuit does not perform any computation rather stores and maps the data from the previous stage to next stage based on the shuffling requirement.

Fig. 1 shows the proposed DMS for a 16-point FFT. The proposed DMS explains how data is processed in the serial FFT processor along with their time indices throughout the stages. The letter ‘*t*’ indicates time indices of all the inputs/outputs. The proposed DMS uses cascaded DR circuits with latency 7 and 2 in stage 1 to re-order the input, while the output data re-ordering begins in stage 4 before the computation of the last BF. The BF operations and the multiplications appear together in all the stages, except last two stages. Thus, the proposed serial architecture requires a total of $\log_2 N - 2$ multipliers. It is important to note that one output of each BF in two consecutive butterflies enters into the same multiplier. One corresponds to the real part while the other corresponds to the imaginary part of a complex BF. The computation of these two butterflies is performed in successive clock cycles as one output from each BF has to be processed through the same multiplier. In stage 3, only DR and BF operations are performed. Note that the most of the arithmetic computations are completed within stage 3. However, there is one more BF computation left to be processed. This allows us to compute all the BF operations in three stages by appending the BF operation of stage 4 at the end of stage 3 using the proposed SSF circuitry.

Unlike the real-valued serial commutator (RSC) FFT [2], the proposed design utilizes half-BF (H-BF) units for computations across all the rows of each stage, except in the last stage. This allows better utilization of H-BF units in the proposed architecture over existing FFT. Moreover, the computational workload on individual H-BF units is also reduced. This can be understood by estimating the H-BF computations performed by individual H-BF units stage-by-stage for the proposed design, i.e. total computational

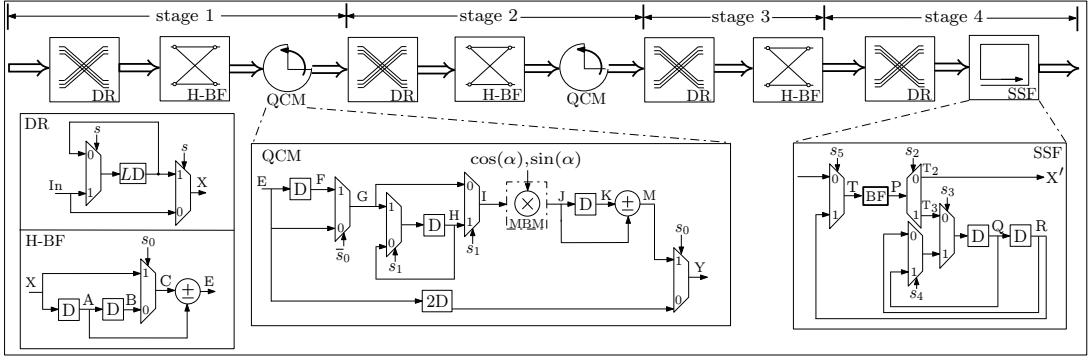


Fig. 2. Proposed 16-point serial pipelined FFT architecture with quarter complex multiplication (QCM) units containing memory-based multiplier (MBM) and select-store-feedback (SSF) unit.

workload on H-BF units by stage-by-stage workload.

$$N + N + N + \dots + 2 = N(\log_2 N - 1) + 2 \quad (1)$$

Similarly, the number of H-BF computations for the existing FFT design [2] can be estimated as $N(\log_2 N + 1) - 6$. Thus, it is clear that the proposed design has an advantage of less computational workload on H-BF units over existing FFT architecture. The percentage reduction in H-BF computations can be expressed as

$$(2N - 8)/[N(\log_2 N + 1) - 6] \times 100\% \quad (2)$$

As per (2), it is clear that reduction in H-BF computations is significant, for instance, the proposed serial FFT offers 30.11% reduction with respect to [2] for 32-points.

III. PROPOSED ARCHITECTURE

Based on the DMS, the proposed serial-pipelined architecture for a 16-point FFT is shown in Fig. 2. It consists of $n = \log_2 N = 4$ stages containing four data re-ordering (DR) circuits, three H-BF units, two quarter complex multiplication (QCM) units which includes the memory-based multiplier (MBM), and a SSF unit. As stated earlier, H-BF, and QCM units are employed together in stage 1 and stage 2, whereas a H-BF unit and a SSF unit are used in stage 3 and stage 4 respectively. Clearly, the design is serial and the H-BF units in $\log_2 N - 1$ stages are always operational in every clock cycle. However, it is not necessary to employ a H-BF unit in the last-stage as it requires one BF computation. This is addressed by the proposed SSF circuitry which first selects the data to be computed, stores it in registers then feeds it back to previous stage. More specifically, as the computations in stage 3 completes, stage 4 BF computation is performed with the same H-BF unit using SSF unit.

A. Data Re-ordering (DR) and Half-Butterfly (H-BF) units:

The data permutations and BF operations shown in the proposed DMS are performed by the DR and H-BF units respectively. The DR and H-BF units employed in the proposed design are identical to [4]. The basic DR unit shown in Fig. 2 consists of a buffer of length ' L ' and two multiplexers to select the stored data or the input. If the

control of multiplexer ' s ' is set to '1', then the sample is passed through the buffer, whereas if it is set to '0', then the input sample at that instant is interchanged with the sample ' L ' clock cycles apart. Unlike H-BF unit, the buffer length of DR unit is different for different stages as observed from Fig. 1. The DR unit in stage 1 of Fig. 2 is used for input reordering, while the DR units from stage 2 to stage $\log_2 N - 1$ are used to shuffle the intermediate data and that in the last stage is used to obtain a normal order output, according to the DMS in Fig. 1. Each stage in the DMS is processed by one stage of the FFT architecture shown in Fig. 2, respectively. To understand the operation of the DR unit, a timing diagram for the cascaded DR units in stage 1 is shown in Table I. The H-BF unit shown in Fig. 2 comprises of two registers, a multiplexer and a real adder. The whole unit computes the addition of BF in one clock cycle and the subtraction of BF in the next clock cycle using a ' s_0 ' signal. The rest of the operation of H-BF unit is explained alongwith QCM in the next sections.

B. Quarter Complex Multiplication (QCM) Unit

In this section, the architectural details of the proposed QCM unit is considered which consists of one real conditional adder, four multiplexers, five registers and one memory-based multiplier (MBM) unit. It is important to note from Fig. 2 that the proposed QCM requires only one register after the real multiplier, unlike [2]. Moreover, there is no need to by-pass the H-BF unit, according to the proposed DMS shown in Fig. 1. Thereby, it reduces the switching activity while transferring the data through or over the MBM unit. As indicated in Fig. 2, the control signal ' s_0 ' decides which data passes on to the multiplier and which data is bypassed. In QCM unit shown in Fig. 2, the first part (E-G) replicates the data, the second part (G-I) which is basically a DR circuit alters the data sequence to be fed into the multiplier while the last part (J-M) performs addition/subtraction of the multiplier outputs to obtain the real and imaginary parts. To understand the complete operation of proposed QCM, a separate timing diagram with H-BF unit in stage 1 is shown in Table I. Note that the letters in the table correspond to various intermediate

TABLE I
TIMING DIAGRAM OF PROPOSED QCM WITH H-BF UNIT AND DR UNIT IN STAGE 1

DR					H-BF					QCM									
Clk	In	s^1	X^1	s^2	X^2	Clk	X	A	B	s_0	E	F	G	H	s_1	I	J	M	Y
0	x_0	1	-	-	-	9	x_0	-	-	-	-	-	-	-	-	-	-	-	-
1	x_1	1	-	-	-	10	x_8	x_0	-	0	x_0^1	-	-	-	-	-	-	-	-
2	x_2	1	-	-	-	11	x_4	x_8	x_0	1	x_8^1	-	x_8^1	-	1	-	-	-	-
3	x_3	1	-	-	-	12	x_{12}	x_4	x_8	0	x_4^1	x_8^1	x_8^1	1	1	x_8^1	cx_8^1	-	x_0^1
4	x_4	1	-	-	-	13	x_2	x_{12}	x_4	1	x_{12}^1	-	x_{12}^1	x_8^1	0	x_{12}^1	dx_{12}^1	$cx_8^1 + dx_{12}^1$	$cx_8^1 + dx_{12}^1$
5	x_5	1	-	-	-	14	x_{10}	x_2	x_{12}	0	x_2^1	x_{12}^1	x_{12}^1	x_8^1	1	x_8^1	dx_8^1	-	x_4^1
6	x_6	1	-	-	-	15	x_6	x_{10}	x_2	1	x_{10}^1	-	x_{10}^1	x_{12}^1	1	x_{12}^1	cx_{12}^1	$cx_{12}^1 - dx_8^1$	$cx_{12}^1 - dx_8^1$
7	x_7	1	x_0	1	-	16	x_{14}	x_6	x_{10}	0	x_6^1	x_{10}^1	x_{10}^1	x_{10}^1	1	x_{10}^1	cx_{10}^1	-	x_2^1
8	x_8	0	x_8	1	-	17	x_1	x_{14}	x_6	1	x_{14}^1	-	x_{14}^1	x_{10}^1	0	x_{14}^1	dx_{14}^1	$cx_{10}^1 - dx_{14}^1$	$cx_{10}^1 - dx_{14}^1$
9	x_9	1	x_2	1	x_0	18	x_9	x_1	x_{14}	0	x_1^1	x_{14}^1	x_{14}^1	x_{10}^1	1	x_{10}^1	dx_{10}^1	-	x_6^1
10	x_{10}	0	x_{10}	1	x_8	19	x_5	x_9	x_1	1	x_9^1	-	x_9^1	x_{14}^1	1	x_{14}^1	cx_{14}^1	$cx_{14}^1 + dx_{10}^1$	$cx_{14}^1 + dx_{10}^1$
11	x_{11}	1	x_4	0	x_4	20	x_{13}	x_5	x_9	0	x_5^1	x_9^1	x_9^1	1	x_9^1	cx_9^1	-	x_1^1	
12	x_{12}	0	x_{12}	0	x_{12}	21	x_3	x_3	x_{13}	1	x_{13}^1	-	x_{13}^1	x_9^1	0	x_{13}^1	dx_{13}^1	$cx_9^1 - dx_{13}^1$	$cx_9^1 - dx_{13}^1$

LEGEND: s^1 : control signal of 1st DR stage, s^2 : control signal of 2nd DR stage, X^1 : output of 1st DR stage, X^2 : output of 2nd DR stage, $c: \cos(\alpha)$, $d: \sin(\alpha)$, x_m^n : output of n^{th} stage with sample index m .

nodes of H-BF and the QCM unit.

TIMING DIAGRAM OF SSF UNIT IN STAGE 4									
Clk	P	s_2	s_3	Q	R	s_4	s_5	T	X'
19	x_0^3	1	0	—	—	—	0	x_2^2	—
20	x_2^3	0	1	x_0^3	—	1	0	x_1^2	—
21	x_1^3	1	0	x_0^3	x_0^3	0	0	x_3^2	x_2^3
22	x_3^3	0	1	x_1^3	x_0^3	0	0	x_{8r}^2	x_3^3
23	x_{8r}^3	0	1	x_1^3	x_0^3	0	0	—	x_{8r}^2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
34	x_{11i}^3	0	1	x_1^3	x_0^3	0	1	x_0^3	x_{11i}^3
35	x_0^4	0	1	x_1^3	x_0^3	0	1	x_1^3	x_0^4
36	x_0^4	0	1	—	—	0	1	—	x_0^4

Fig. 3. Timing diagram of SSF unit in stage 4, where x_n^m denotes output of m^{th} stage with sample index n .

C. Select-Store-Feedback (SSF) unit

The proposed SSF unit as shown in Fig. 2 includes one de-multiplexer, three multiplexers and two registers. As the name suggests, it selects and stores the last two samples to be computed in stage 4, and feeds them back until the computation of H-BF unit in stage 3 is over, as shown in Fig. 1. It is then unloaded to the H-BF unit of stage 3 for the last BF computation. For example, consider the proposed DMS scheme for a 16-point RFFT shown in Fig. 1. The samples with time indices 28 and 30 are computed in stage 4, therefore, it is necessary to store them till the computation of stage 3 is over. The control signal s_2 of de-multiplexer is set to 1 when the computation of sample with time-index 28 is over in stage 3, and is stored in the first register by setting $s_3 = 0$. However, it is required to skip the sample with time-index 29 (in stage 4) as it is not needed in BF computation of stage 4. Hence, it is passed via T_2 line of de-multiplexer like other BF computations of stage 3. The control signals s_4 and s_5 are used to store sample with time-index 30 and utilize the H-BF unit of stage 3 in the end, respectively. The operation of the proposed SSF unit is explained in Fig. 3.

D. Memory based multiplier (MBM)

The principle involved in memory-based multiplication is shown in Fig. 4(a). It contains a memory unit and external

combinational logic, where memory stores the partial products and combinational logic decodes the stored contents of memory. For the simplicity of discussion, the inputs (X) and twiddle coefficients (A) are assumed to be W -bit and B -bit unsigned binary numbers respectively, that is, $X = x_{W-1}x_{W-2}\dots x_0$ and $A = a_{B-1}a_{B-2}\dots a_0$. Note that sign-magnitude or two's complement can be described in the same manner except most-significant bit (MSB) which would be used as sign-bit. In multiplication, if input is assumed to be the multiplier and coefficient to be the multiplicand, then the total number of partial products to be stored would be 2^W . However, this way of storing the partial products in memory is cost inefficient since the complexity of memory grows exponentially with wordlength of input, while the cost of combinational logic is just an address decoder. This is discussed in [9], where single memory unit is used as a look-up-table containing partial products corresponding to all possible values of input X . As a result, a bit mapping scheme (BMS) is suggested which basically encodes W -bits to $(W-1)$ -bits at slight increase in combinational logic while reducing the memory complexity by half. In other words, only half number of partial products are required to be stored while the other half partial products are decoded through relatively more complex external combinational logic. This is depicted in the truth table of Fig. 5 where memory has eight address locations addressed with the outputs of BMS unit. The proposed work stores only odd partial products as it is difficult to generate partial products on hardware [9]. While the remaining partial products can be obtained through external BMS unit by reading the memory content followed by extra combinational logic in terms of shifting. So far, a memory with combinational logic can be used for multiplication of an W -bit input with a B -bit coefficient as:

- A memory unit with word size $(W+B)$ -bits is employed to store odd partial products while even partial products are produced through the left shifts on the stored contents.
- To decode the odd partial products, they are mapped to address space of memory using

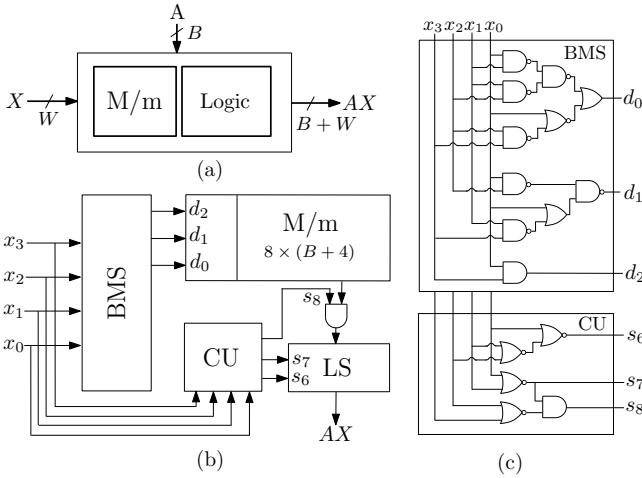


Fig. 4. (a) Concept of memory based multiplication (b) Block schematic of memory based multiplication with input operand of size 4-bit, $X = x_3x_2x_1x_0$ and twiddle coefficients are of B -bit (c) Gate level implementation of bit-mapping scheme (BMS) and control unit (CU). LS: Logarithmic shifter or barrel shifter.

the outputs $d_{W-2}d_{W-3}\dots d_0$ of BMS, and the shifts for the logarithmic-shifter with control bits $s_6s_7\dots s_{\lceil \log_2 W \rceil + 1 + 6}$, where $\lceil \cdot \rceil$ is ceil function.

- Note that when all the bits of input are zero, the product value must be zero which is obtained by ANDing control bit s_8 with output of memory.

For clarity, let us consider 4-bit multiplier with architectural details of BMS unit, logarithmic-shifter (LS) and a control unit (CU). The boolean expression of d_2, d_1 , and d_0 can be obtained by simplifying the logic shown in Fig. 4(c) using Karnaugh map. Accordingly, $d_0 = \overline{(x_0x_1)} \cdot \overline{(x_1x_2)} \cdot (x_0 + \overline{(x_2x_3)})$, $d_1 = \overline{(x_0x_2)} \cdot (x_0 + (x_1x_3))$, $d_2 = x_0x_3$ respectively. Note that the boxes shaded in gray colour are pre-computed partial products which are stored in memory. It is clear from Fig. 5 that the number of left-shifts required on the stored-word is three, therefore, two-stage logarithmic-shifter is sufficient. The reset function is implemented by a AND-cell with inputs as control signal s_2 and output of memory, as shown in Fig. 4(b) using an active-low RESET.

IV. PERFORMANCE COMPARISON

In this section, the computational complexities of the proposed and existing serial-pipelined FFT designs are discussed. Next, the performance of the proposed design along with the existing designs are evaluated and compared through ASIC synthesis and FPGA implementation in terms of area, power, minimum-clock period (MCP), area-delay product (ADP), sliced look-up table (SLUT) and flip-flops (FF). For simplicity, the designs in [2], [4], [10], [11] are referred as FFT₀, FFT₁, FFT_{2,3} and FFT₄ respectively.

A. Computational Complexities

The computational complexities of the proposed and existing designs [2], [4], [10], [11] in terms of real multipliers, real adders and registers are listed in Table II

input $x_3x_2x_1x_0$	BMS $d_2d_1d_0$	M/m	LS Shifts	CU s_7s_6
0 0 0 0	s_8	—	—	—
0 0 0 1	0 0 0	A	<< 0	0 0
0 0 1 0	0 0 0	2A	<< 1	0 1
0 0 1 1	0 0 1	3A	<< 0	0 0
0 1 0 0	0 0 0	4A	<< 2	1 0
0 1 0 1	0 1 0	5A	<< 0	0 0
0 1 1 0	0 0 1	6A	<< 1	0 1
0 1 1 1	0 1 1	7A	<< 0	0 0
1 0 0 0	0 0 0	8A	<< 3	1 1
1 0 0 1	1 0 0	9A	—	—
1 0 1 0	0 1 0	10A	<< 1	0 1
1 0 1 1	1 0 1	11A	—	—
1 1 0 0	0 0 1	12A	<< 2	1 0
1 1 0 1	1 1 0	13A	—	—
1 1 1 0	0 1 1	14A	<< 1	0 1
1 1 1 1	1 1 1	15A	—	—

M/m stores the partial product corresponding to shaded box

Fig. 5. Memory contents and partial products for input wordlength $W = 4$, s_7 and s_6 control signals for logarithmic shifter (LS). BMS: bit-mapping scheme, CU: control unit.

alongwith an example of hardware complexity for a 1024-point FFT. All of the architectures are serial in nature, therefore, they process one sample per clock cycle. As a result, its throughput and latency is one sample per clock cycle and N clock cycles respectively. Note the designs referred as FFT₁ and FFT_{2,3,4} process complex-valued and real-valued data respectively. However, the implementation costs of all the designs are relatively higher due to complex multipliers and BF computations. As listed in Table II, the cost of adders for the proposed design is always less than all the existing designs. It is basically due to absence of adder in the last stage of FFT computation. In addition, the data computed in stage 4 is selected, stored and fed back to the stage 3 H-BF unit with the proposed SSF unit. As a result, it has an advantage of less usage of hardware resources with respect to existing designs. The FFT_{2,3} designs support continuous-flow operation, but involve redundant operations when applied to real-valued FFT computations. Further, real-valued FFT architectures reduce the memory requirements by a factor of two over complex-valued FFT architectures. On the other hand, the architecture in [4] is based on CORDIC whose output is dependent on iterations inducing error and latency. This problem is addressed with the proposed memory-based multiplier for the FFT computation. The proposed MBM unit consists of memory unit and combinational logic, and results in computation savings for realizing multiplier. Interestingly, the savings obtained for the case of proposed design have direct impact on the dynamic power consumption, which would increase with the number of FFT points.

B. Implementation Results

The proposed and existing designs are coded in Verilog for 1024-point FFT with wordlength of inputs and coefficients considered as 8-bit and 16-bit respectively. Subsequently, application specific integrated circuit (ASIC)

TABLE II
HARDWARE COMPLEXITIES OF VARIOUS SERIAL PIPELINED FFT

Design	N			N=1024	
	MULT	ADD	NOR	MULT	ADD
Complex-valued data					
FFT [†] ₀ [4]	0	$6\log_2 N - 2$	-	0	232
FFT [‡] ₁ [12]	$4\log_2 N - 8$	$6\log_2 N - 4$	$2N$	32	56
FFT [‡] ₂ [10]	$2\log_2 N - 4$	$3\log_2 N - 2$	$\approx 2N$	16	28
Real-valued data					
FFT [‡] ₃ [11]	$4\log_2 N - 12$	$6\log_2 N - 10$	$5N/4 - 2$	28	50
FFT [‡] ₄ [11]	$4\log_2 N - 12$	$4\log_2 N - 6$	$3N/2 - 5$	28	34
FFT [‡] ₅ [2]	$\log_2 N - 2$	$2\log_2 N - 2$	$N + 9\log_2 N - 19$	8	18
Proposed	0	$2\log_2 N - 3$	$N + 7\log_2 N - 18$	0	18

LEGEND: †: CORDIC-based, ‡: Multiplier-based, MULT: Multipliers, ADD: Adders, NOR: Number of registers. Note the throughput and latency of all the designs are one sample-per-clock cycle and N clock cycles respectively.

synthesis is performed by Cadence 14.1 RTL compiler using TSMC 90 nm CMOS technology. The obtained synthesis results are listed in terms of area, power, minimum-clock period (MCP) and area-delay product (ADP) as shown in Table III. As expected, the proposed design occupies significantly less area and power as compared to existing designs. It is because the proposed design does not include physical multiplier as in existing designs and rather realizes it with half-sized memory and combinational logic. In addition, the proposed design offers lower MCP over the existing design [2] due to lower memory access time. The proposed architecture for 1024-point occupies nearly 31.54% less area, 30.13% less power and 33.56% less ADP than [2]. Clearly, the savings obtain with respect to [2] are significant. Further savings in power consumption is possible by employing clock and power gatings as they do not depend on the processing of previous stages. The proposed and existing designs are also implemented on a Xilinx ZYNQ field programmable gate array (FPGA) device (XC7Z020-1CLG84C) for 1024-point RFFT. The logic utilization is obtained in terms of slice LUTs (SLUT) and flip-flops (FF) by setting the system clock at 50 MHz, as listed in Table II. From the implementation results, it is clear that the proposed design for 1024-point FFT has about 27.11% less SLUT and 28.37% less FF as compared to the design in [2].

V. CONCLUSION

In this paper, a high-performance architecture for serial pipelined FFT based on new data mapping scheme is presented. The proposed data management scheme allows reduction in the utilization of BF units leading to performance enhancement in area and power. It performs FFT computations in $\log_2 N - 1$ stages followed by a select-store-feedback stage. The proposed architecture employs memory units and combinational logic in place of multipliers. By doing so, further improvements in area and power has been achieved since memory stores only half number of partial products. ASIC synthesis and FPGA implementation of 1024-point real-valued FFT showed that the proposed design has significant improvement in performance, viz. 31.54% less area, 30.13% less power, 33.56% less ADP, 27.11%

TABLE III
ASIC SYNTHESIS USING 90 NM CMOS LIBRARY FOR 1024-POINTS SERIAL PIPELINED RFFT ARCHITECTURES

Design	Area (mm ²)	Power (mW)	MCP (ns)	ADP (mm ² × ns)	SLUT (×1000)	FF (×1000)
FFT ₀ [4]	1.614	4.86	7.65	12.347	61.29	55.32
FFT ₁ [12]	2.216	6.17	9.58	21.229	83.77	68.22
FFT ₂ [10]	1.255	3.75	10.11	12.688	51.34	47.21
FFT ₃ [11]	1.572	5.04	7.15	11.239	59.13	52.56
FFT ₄ [11]	1.133	4.12	6.56	7.432	46.84	42.66
FFT ₅ [2]	1.043	3.12	6.14	6.404	42.89	39.87
Proposed	0.714	2.18	5.96	4.255	31.26	28.56

LEGEND: MCP: minimum-clock period, ADP: area-delay product, SLUT: sliced look-up table, FF: flip-flop.

less SLUT and 28.37% less FF as compared to the latest reported scheme.

VI. ACKNOWLEDGEMENT

This work was supported by Special Manpower Development Programme for Chip to System Design (SMDP-C2SD) sponsored by the Ministry of Electronics & Information Technology (MeitY), Govt. of India.

REFERENCES

- [1] Z.-G. Ma, X.-B. Yin, and F. Yu, "A novel memory-based FFT architecture for real-valued signals based on a radix-2 decimation-in-frequency algorithm," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 9, pp. 876–880, 2015.
- [2] M. Garrido, N. K. Unnikrishnan, and K. K. Parhi, "A serial commutator fast Fourier transform architecture for real-valued signals," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2017.
- [3] Z. Wang, X. Liu, B. He, and F. Yu, "A combined SDC-SDF architecture for normal I/O pipelined radix-2 FFT," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 5, pp. 973–977, 2015.
- [4] S. S. Abdullah, H. Nam, M. McDermot, and J. A. Abraham, "A high throughput FFT processor with no multipliers," in *Computer Design, 2009. ICCD 2009. IEEE International Conference on*. IEEE, 2009, pp. 485–490.
- [5] M. Garrido, R. Andersson, F. Qureshi, and O. Gustafsson, "Multiplierless unity-gain SDF FFTs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 9, pp. 3003–3007, 2016.
- [6] M. T. Khan and R. A. Shaik, "Optimal complexity architectures for pipelined distributed arithmetic-based LMS adaptive filter," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2018.
- [7] M. Ayinala and K. K. Parhi, "FFT architectures for real-valued signals based on radix-2³ and radix-2⁴ algorithms," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 9, pp. 2422–2430, 2013.
- [8] W. M. Arden, "The international technology roadmap for semiconductors-perspectives and challenges for the next 15 years," *Current Opinion in Solid State and Materials Science*, vol. 6, no. 5, pp. 371–377, 2002.
- [9] P. K. Meher, "New approach to LUT implementation and accumulation for memory-based multiplication," in *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on*. IEEE, 2009, pp. 453–456.
- [10] M. Garrido, S.-J. Huang, S.-G. Chen, and O. Gustafsson, "The serial commutator FFT," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 10, pp. 974–978, 2016.
- [11] A. Chinnapalanichamy and K. K. Parhi, "Serial and interleaved architectures for computing real FFT," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1066–1070.
- [12] S. He and M. Torkelson, "Design and implementation of a 1024-point pipeline FFT processor," in *Custom Integrated Circuits Conference, 1998. Proceedings of the IEEE 1998*. IEEE, 1998, pp. 131–134.

Efficient Methods for Estimating Sinusoidal Frequencies Using Line Spectral Pairs

P. Vishnu and C.S. Ramalingam

Department of Electrical Engineering, IIT Madras, Chennai-600036

Email: {ee12d038,csr}@ee.iitm.ac.in

Abstract—The maximum likelihood (ML) method of estimating the frequencies of p sinusoids in the presence of AWGN is computationally very costly because of the dimensionality of the error surface; the advantage is that the ML method has the lowest threshold among all known practical estimators. We propose a low complexity method using Line Spectral Pairs (LSPs), where the LSPs are derived from an estimated $A(z)$ obtained using Multiple Signal Classification (MUSIC) method. The proposed method evaluates the likelihood function at significantly fewer number of points—at most ${}^5 p C_p$ —for getting the estimates. Furthermore, no iterative finer search is required. Nevertheless, the proposed method’s threshold is comparable to that of ML when tested using the well-known two-sinusoids example; similar performance was observed in the case of three sinusoids. Further improvements were observed when the beamformer function was used for detecting and removing outliers. For the two-sinusoid case, outlier removal resulted in a threshold that was lower than that of ML by as much as 9 dB (3 $\pi/2$ case). We also present results for a direction of arrival (DOA) estimation example that results in the same threshold as that of ML.

I. INTRODUCTION

In this paper we address the classical problem of estimating frequencies of multiple sinusoids. It has applications in several fields such as radar, sonar, radio astronomy, etc. [1]–[3]. For example, the problem of estimating direction of arrival (DOA) of signals on a uniform linear array can be modeled as a sinusoidal frequency estimation problem [4] [3, Ch. 6].

The observed signal $x[n]$ is modeled as a sum of p complex sinusoids corrupted by additive white Gaussian noise $w[n]$ with variance σ_w^2 . That is,

$$x[n] = \sum_{k=1}^p \alpha_k e^{j2\pi f_k n} + w[n] \quad n = 0, 1, \dots, N-1 \quad (1)$$

where $\alpha_k (= |\alpha_k| e^{j\phi_k})$ ’s and f_k ’s are the unknowns. The problem of estimating f_k ’s is difficult due to the nonlinear dependence of $x[n]$ on the f_k ’s. Once f_k ’s are estimated, estimation of α_k ’s is straightforward [2]. In this paper we assume p is known; if p is unknown, one can use a detection scheme to estimate it (e.g., [5]).

The periodogram method [1] is computationally the least expensive but is limited by low frequency resolution: it can resolve two signals only if $|f_k - f_l| > 1/N$ [2], [6]. The method of maximum likelihood (ML) estimation has the ability to resolve two closely spaced sinusoids, which comes at a significant computational cost. Linear prediction based

methods such as MUSIC method [7], [8] are suboptimal high resolution methods with moderate computational complexity.

All practical estimators exhibit a sudden sharp increase in the variance of the estimates as the SNR becomes lower. The point at which this happens is termed as the *threshold SNR*. The ML estimator has the lowest threshold among all known practical estimators. However, since it involves a p -dimensional coarse grid search, followed typically by a gradient descent algorithm, it ranks among the most computationally expensive methods [2]. The FastML method proposed in [9] reduces the ML’s complexity to only p one-dimensional searches. However, it gave very poor results for the three-sinusoid example considered in this paper. In the context of DOA estimation, Shaghaghi and Vorobyov [10] use the ML objective for selecting that subset of the roots provided by root-MUSIC that are most likely to be signal roots. They called this as the “root-swap method”.

In this paper we propose a low-complexity method, wherein we evaluate the ML function at a small set of points; these points are obtained using Line Spectral Pairs (LSP) [11], [12]. This method achieves the threshold SNR of the ML method for most phase differences in the well-known two-sinusoid example [2], [13]. For removing outliers, hypothesis tests based on the conventional beamformer at the candidate frequencies is used [14]–[16]. Outlier removal further improves the threshold SNR.

In Section II-A ML and root-MUSIC are reviewed; LSPs are reviewed in Section-II-B. The low complexity methods that exploit the information given by the LSPs are given in Section-III. Simulation results are given in Section-IV.

II. REVIEW

A. ML and Root-MUSIC

Let $\mathbf{x} = (x[0] \ x[1] \ \dots \ x[N-1])^T$ be the observed data and $\mathbf{e}_i = [1 \ e^{j2\pi f_i} \ \dots \ e^{j2\pi(N-1)f_i}]^T$. In ML estimation, the frequency estimates are obtained by maximizing the following objective function $L(\mathbf{f})$ [2]:

$$L(\mathbf{f}) = \mathbf{x}^H \mathbf{S} (\mathbf{S}^H \mathbf{S})^{-1} \mathbf{S}^H \mathbf{x} \quad (2)$$

where $\mathbf{S} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \dots \ \mathbf{e}_p]$. Direct implementation involves a p -dimensional coarse search on a frequency grid followed by finer search using gradient descent based algorithms [2].

In root-MUSIC [3], [7], the annihilating filter $A(z) = 1 + \sum_{k=1}^M a_k z^{-k}$ is estimated using the noisy subspace eigenvectors

of the autocorrelation matrix [3]. If $A(z)$ has p roots on the unit circle at $\omega_k = 2\pi f_k$ (the “signal roots”) [2], [3], then it annihilates signal component of $x[n]$ due to $A(e^{j\omega_k}) = 0$ [3]. In practice, the estimated $A(z)$ will have signal roots not on the unit-circle but close to it. The value of M is usually chosen to be greater than p for noise robustness; $M = 3N/4$ was recommended in [13] based on empirical observations. Ideally, the remaining $M - p$ roots (the “extraneous roots”) should be well inside the unit circle. In this method, the angles of the p roots of $A(z)$ that are closest to the unit circle [3] are taken as the frequency estimates. Estimating the autocorrelation matrix by the method of forward-backward averaging usually yields the best results [3].

B. Line Spectral Pairs

The LSP polynomials $P(z)$ and $Q(z)$ are defined as $P(z) = A(z) + z^{-(M+1)}A^*(1/z^*)$ and $Q(z) = A(z) - z^{-(M+1)}A^*(1/z^*)$. From the definition it should be clear that $P(z)$ is conjugate symmetric, whereas $Q(z)$ is conjugate anti-symmetric. It is well-known that if $A(z)$ is minimum phase, then the roots of $P(z)$ and $Q(z)$ lie on the unit circle and interlaced [11], [12], [17]–[19]. These properties holds true for complex LSP polynomials also (proof given in Appendix A).

Only in the noiseless case is the estimated $A(z)$ guaranteed to have exactly p zeros on the unit circle at the signal frequencies, i.e., $A(e^{j\omega_k}) = 0$ for $k = 1, 2, \dots, p$. In the noisy case, if the estimated $A(z)$ is not minimum phase, it is made so by reflecting the offending roots about the unit circle.

It can be shown that $4|A(e^{j\omega})|^2 = |P(e^{j\omega})|^2 + |Q(e^{j\omega})|^2$, which is a consequence of $P(e^{j\omega})Q^*(e^{j\omega}) + P^*(e^{j\omega})Q(e^{j\omega}) = 0$ (see Appendix B). Hence it follows that $P(e^{j\omega_k}) = Q(e^{j\omega_k}) = 0$. That is, the zeros of both $P(z)$ and $Q(z)$ coincide and correspond to the signal zeros in the noise-free case.

It is well-known that a root close to the unit circle gives rise to an LSP whose elements are close to each other [17], [19]. This can also be seen from the following equation, which is the extension to the complex-valued case of the result that is given in [20] for the real-valued counterpart (proof given in Appendix B):

$$|A(e^{j\omega})|^2 = \frac{\prod_{k=1}^{M+1} \sin^2 \left(\frac{\omega - \omega_{P_k}}{2} \right) + \prod_{l=1}^{M+1} \sin^2 \left(\frac{\omega - \omega_{Q_l}}{2} \right)}{4^{-M}} \quad (3)$$

where ω_{P_k} and ω_{Q_l} are the corresponding LSP angles of $P(z)$ and $Q(z)$. Since $|A(e^{j\omega})|^2 \approx 0$ near a signal root, there must exist k and l for which $(\omega_{P_k}, \omega_{Q_l})$ is a close LSP (because both $|\omega_{P_k} - \omega|$ and $|\omega_{Q_l} - \omega|$ must be small). We exploit the above relationship between a root close to the unit circle and its corresponding close LSP in our proposed method.

Let z_i for $i = 1, 2, \dots, M$ be the zeros of $A(z)$. Let $\theta_k \in [0, 2\pi)$ for $k = 1, 2, \dots, 2M + 2$ represent the angles of the roots of $P(z)$ and $Q(z)$ in ascending order (successive root angles in the counter-clockwise direction). Define an LSP as the pair of angles of the form (θ_i, θ_{i+1}) , where i is $((i-1)$

$\mod 2M + 2) + 1$ (which helps i to wrap circularly). Let (θ_l, θ_{l+1}) be an LSP “near a signal root”; we term this pair as signal LSP. The measure of nearness is defined in the next section.

III. LOW COMPLEXITY METHODS USING LSPS

In the ML method we estimate $\hat{\mathbf{f}}$ by maximizing $L(\mathbf{f})$ (Eq. (2)). Since $L(\mathbf{f})$ typically has local maxima, a coarse grid search is employed in the associated p -dimensional space, which is then followed by a gradient descent procedure. This is computationally intensive. We aim to reduce the computational burden without compromising on the threshold SNR that the ML method gives. This is achieved by evaluating $L(\mathbf{f})$ on a small set of frequency points located in the neighborhood of the estimates given by root-MUSIC. These set of points are derived from the LSPs of $A(z)$.

Using the LSPs to obtain search grid points is motivated by the examples given in Fig. 1. The high SNR case (20 dB) is shown in Fig. 1(a), with root-MUSIC estimating the signal frequencies correctly. The corresponding signal LSPs are $(\theta_{13}, \theta_{14})$ and $(\theta_{15}, \theta_{16})$. Observe that θ_{12} is far apart from θ_{13} ; similarly, θ_{14} is far away from θ_{15} . On the other hand, for the low SNR case (5 dB), it can be seen from Fig. 1(b) that one of the signal frequencies as estimated by root-MUSIC has a large error. However, if we examine the correctly estimated signal LSP, we find that its neighbors are closer, i.e., θ_{14} is close to θ_{15} and θ_{17} is close to θ_{16} . In such cases, another signal frequency can be found with high probability in this neighborhood. For practical implementation, we have chosen 5 points around a signal LSP. For the p -component case, the total number of points chosen around p signal LSPs will be at most $5p$.

The computational steps for choosing the $5p$ points are:

- 1) Obtain initial estimates \check{f}_k , $k = 1, 2, \dots, p$ using root-MUSIC
- 2) Initialize $i = 1$
- 3) Find LSPs (θ_l, θ_{l+1}) near to \check{f}_i using

$$l = \operatorname{argmin}_k |2\pi\check{f}_i - \theta_k| + |2\pi\check{f}_i - \theta_{k+1}|$$

- 4) The search points in the neighborhood of (θ_l, θ_{l+1}) is given by

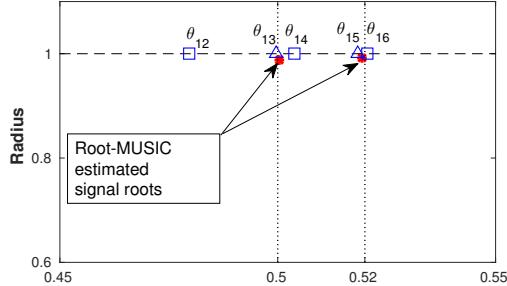
$$\Omega_i = \left\{ \frac{\theta_{l-1} + \theta_l}{2}, \theta_l, \frac{\theta_l + \theta_{l+1}}{2}, \theta_{l+1}, \frac{\theta_{l+1} + \theta_{l+2}}{2} \right\}$$

- 5) $i \leftarrow i + 1$ and repeat steps 3–4 for $i \leq p$
- 6) The overall search set is

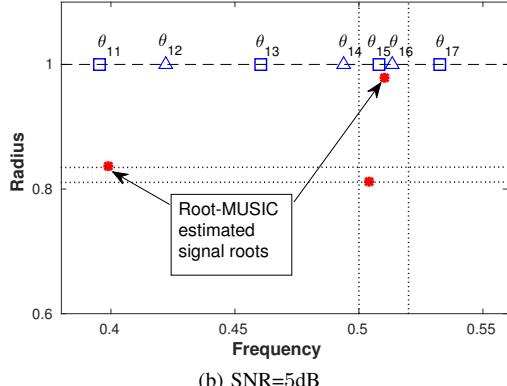
$$\Omega = \Omega_1 \cup \Omega_2 \cup \dots \cup \Omega_p$$

Note that Ω is a set with a maximum of $5p$ points ($5p$ if there is no duplication).

The above can be interpreted as a coarse grid search that enables us to reduce the number of outliers at low SNR when there are closely spaced components. As the SNR increases, the signal roots move closer to the unit circle and closeness of signal LSPs also increases. Hence, θ_l , θ_{l+1} and its midpoint



(a) SNR=20dB



(b) SNR=5dB

Fig. 1. LSPs and roots of $A(z)$ are plotted for two sinusoids case with $f_1 = 0.5$, $f_2 = 0.52$, $\phi_1 - \phi_2 = 0$ (a) High SNR example (20 dB) where Root-MUSIC estimates signal roots correctly. (b) Low SNR (5 dB) example where Root-MUSIC estimates signal roots wrongly.

become finely spaced. Thus the above can be interpreted as finer grid search, which improves the accuracy of the estimates at high SNR.

Once the $5p$ points have been chosen, we propose two methods to estimate the signal frequencies.

A. Method 1

In this method $\hat{\mathbf{f}}$ is chosen to be the \mathbf{f} that maximizes the likelihood function $L(\mathbf{f})$. That is,

$$\hat{\mathbf{f}} = \operatorname{argmax}_{\mathbf{f}} \{L(\mathbf{f}) \mid 2\pi\mathbf{f} \subset \Omega\} \quad (4)$$

Since Ω contains at most $5p$ frequencies, and since \mathbf{f} is a p -dimensional vector, one has to carry out at most ${}^{5p}C_p$ number of evaluations of $L(\mathbf{f})$, at the end of which we identify $\hat{\mathbf{f}}$.

B. Method 2

The performance of the Method 1 can be improved if we can identify and remove the outliers present in the search space. In the context of DOA estimation it has been shown that the outliers can be identified by carrying out hypothesis tests on the conventional beamformer response evaluated at the candidate frequencies [14]–[16].

The beamformer response is given by

$$P(\omega) = \mathbf{e}(\omega)^H \hat{\mathbf{R}} \mathbf{e}(\omega) \quad (5)$$

where $\hat{\mathbf{R}}$ is the estimated autocorrelation matrix and $\mathbf{e}(\omega) = [1 \ e^{j\omega} \ e^{j2\omega} \ \dots \ e^{j(M)\omega}]^T$. Let Φ be the pre-estimated angular

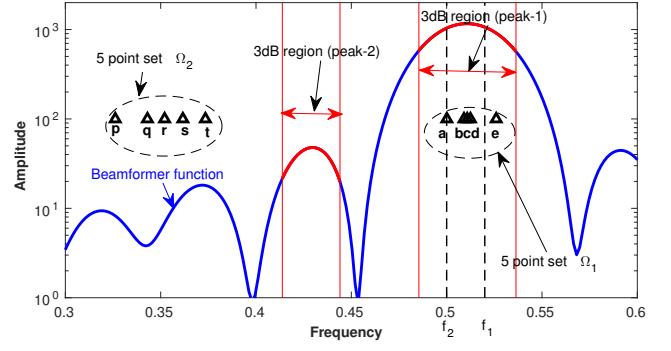


Fig. 2. Beamformer response and positions of $5p$ points for the two sinusoids example (SNR=12dB, $\phi_1 - \phi_2 = 3\pi/2$). Method 1 gives the estimates as r and c . The outliers p, q, r, s and t are removed using the beamformer hypothesis test. The final estimates obtained using Method 2 are a and e .

sectors using the beamformer response [14], [15]. It is given by

$$\Phi = [\phi_1^{\max} - \phi_1^{\text{left}}, \phi_1^{\max} + \phi_1^{\text{right}}] \cup \dots \cup [\phi_p^{\max} - \phi_p^{\text{left}}, \phi_p^{\max} + \phi_p^{\text{right}}] \quad (6)$$

where $\phi_i^{\max}, i = 1, \dots, p$ are the top p peaks of the beamformer response, with ϕ_i^{left} and ϕ_i^{right} being the left and right boundaries of the i -th subinterval. They can be chosen as the 3 dB points. The hypothesis to be tested is whether or not all the frequency estimates are localized in Φ .

In the proposed method, the outliers in the search space Ω can be removed using the above hypothesis test. Let $\Omega_H \subseteq \Omega$ be the new search space in which all the elements of Ω_H succeed the hypothesis test. Then, Ω_H is used in place of Ω in Eq. (4) for finding the final frequency estimates.

To illustrate outlier removal, consider the two-sinusoid example with $\phi_1 - \phi_2 = 3\pi/2$ at an SNR of 12 dB. The threshold of the ML estimator for this case is 15 dB. Thus this SNR is well below threshold of ML. Fig. 2 shows the main portions of beamformer response function $P(\omega)$. Also shown are the $5p = 10$ search points, with $\Omega_1 = \{a, b, c, d, e\}$ and $\Omega_2 = \{p, q, r, s, t\}$. Using Method 1, the estimated frequencies are frequencies r and c . Interestingly, the ML estimates are also close to r and c . Clearly, the estimate r has a large error, i.e., it is an outlier. This outlier can be eliminated using the hypothesis test. To identify the outliers, the frequency regions that fall within the 3 dB range of the top $p = 2$ peaks are considered. This is the set Φ . Since $\{p, q, r, s, t\} \notin \Phi$, we can eliminate these points from consideration when evaluating $L(\mathbf{f})$. As a result, $L(\mathbf{f})$ is evaluated only at $\{a, b, c, d, e\}$. This leads to a and e being identified as the final frequency estimates.

IV. SIMULATION RESULTS

To showcase the performance improvement over the existing methods we give simulation results for the well-known two-sinusoids example. To show that the improvement using proposed method is not restricted to the two sinusoids case, we have given results for a three sinusoids example; we have also

applied our methods to the DOA example of Shaghaghi and Vorobyov [10]. 50,000 trials were used in all the simulations. For the sinusoids examples, we have used $N = 25$ and $M = 18$. We plot MSE versus SNR, where the MSE is defined as the sum of mean-squared errors of $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_p$.

A. Two sinusoids

We use the well-known two sinusoids example [13]: $|\alpha_1| = |\alpha_2| = 1$, $f_1 = 0.52$, $f_2 = 0.5$. The MSE plots for the different methods are shown in Fig. 3 for $\phi_1 - \phi_2 = 0$. Method 1 is significantly better than root-MUSIC; Method 2 is similar to that of Method 1 and has the same threshold as ML. The bias of proposed methods above threshold are comparable to that of root-MUSIC, with values ranging from 10^{-4} to 10^{-5} . For the $\phi_1 - \phi_2 = 3\pi/2$ case (Fig. 4), Method 1 is better than root-MUSIC but poorer than ML in terms of threshold. However, the removal of outliers in Method 2 results in a threshold that is *lower than that of ML*. The improvement is as much as 9 dB.

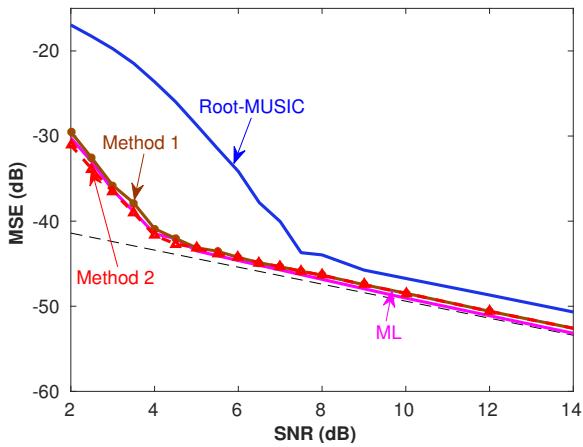


Fig. 3. Estimation performance of different methods for two sinusoids example with $\phi_1 - \phi_2 = 0$. Method 1 achieves threshold SNR of ML method. Performance of Method 2 is similar to Method 1.

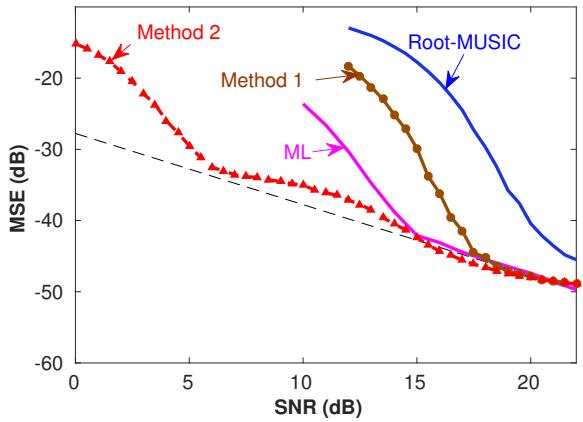


Fig. 4. For the $\phi_1 - \phi_2 = 3\pi/2$ case, the threshold SNR of Method 1 is slightly poorer than that of ML method. Method 2 has a significantly lower threshold of 6 dB, which is better than that of the ML method by as much as 9 dB!

The threshold SNRs for various phase differences $\phi_1 - \phi_2$ in the 0 to 2π in steps of $\pi/4$ is plotted in Fig. 5. It shows that Method 1 nearly or exactly matches the threshold SNR of ML in 6 out of 8 cases. Method 2 is either same or better than Method 1 in every single case. Furthermore, Method 2 has better threshold SNR than ML whenever $\phi_1 - \phi_2 > \pi$.

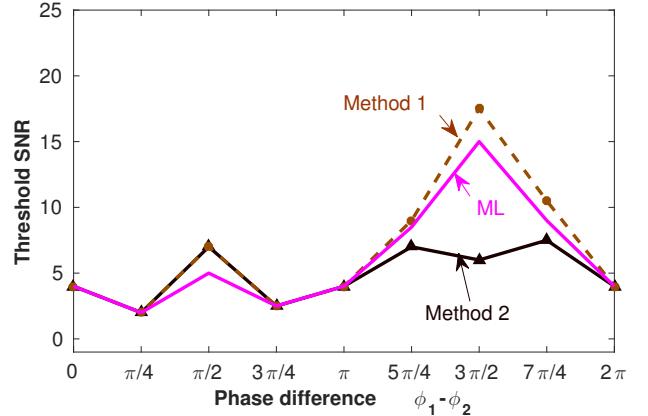


Fig. 5. Comparison of threshold SNR for two sinusoids example. Method 1 achieves the threshold SNR of ML method for many values of $(\phi_1 - \phi_2)$. Method 2 is better than the ML method for $(\phi_1 - \phi_2) > \pi$.

To get an idea of the computational savings, the ML function was evaluated at a set of 124,750 point prior to the gradient descent step (500 points per dimension, but only half of them are needed since it is enough if we consider only the region $f_1 > f_2$). On the other hand, the value of ${}^{5p}C_p$ for $p = 2$ is merely 45, which is significantly lower.

B. Three sinusoids

1) *Three sinusoids with one far away and two closely spaced frequencies:* To the two-sinusoid example we added one more sinusoid that is far away, i.e., $f_3 = 0.3$ with $\phi_1 = 0, \phi_2 = \pi/4$ and $\phi_3 = 0$. The MSE curves plotted in Fig. 6 reveal that Method 1 is significantly better than root-MUSIC method and has the same threshold SNR as that of

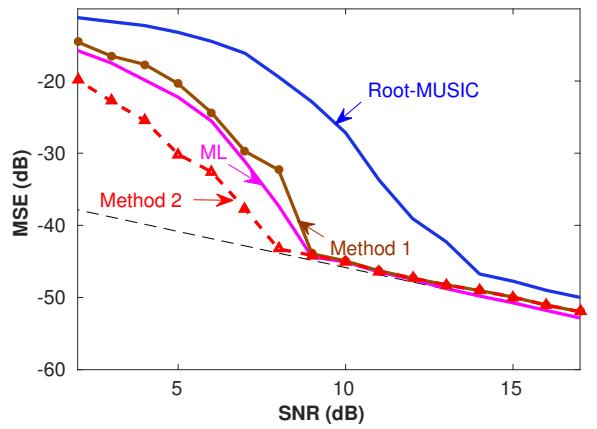


Fig. 6. Estimation performance when a third far away sinusoid is added $f_3 = 0.3$. Method 2 has a lower threshold than that of ML.

ML method. For this example also, Method 2 has a threshold that is lower than that of ML.

2) *Three closely spaced sinusoidal frequencies:* In this example the third sinusoid is also closely spaced with $f_3 = 0.48$ and $\phi_1 = 0, \phi_2 = \pi/4$ and $\phi_3 = 0$. The results for this case are shown in Fig. 7. Method 1 is much better than root-MUSIC and almost comparable to ML. But once again Method 2 performs the best, and significantly outperforms ML, having a threshold that is 9 dB lower.

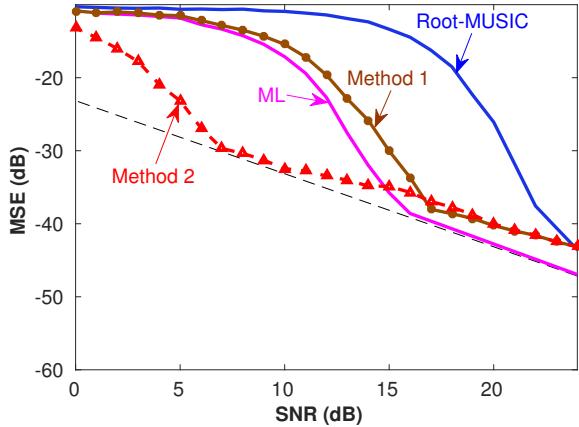


Fig. 7. Estimation performance for three sinusoid case with all closely spaced frequencies, where $f_1 = 0.52, f_2 = 0.5, f_3 = 0.48$. Method 2 has a threshold that is lower than 9 dB when compared to ML.

C. DOA estimation

In DOA estimation, the data consist of L snapshots of length M ($M \times L$) matrix; the autocorrelation matrix of size $M \times M$ is obtained using the forward-backward averaging method from L snapshots. To showcase the performance improvements, we have used the same test example given in Shaghaghi and Vorobyov [10]. In this example $p = 2$ sources are impinging on an array of $M = 10$ antenna elements from directions $\psi_1 = 35^\circ(\pi/180)$ and $\psi_2 = 37^\circ(\pi/180)$. The inter-element spacing is set to $d = \lambda/2$ and the number of snapshots $L = 10$ is used. The source distribution is $\mathcal{N}(\mathbf{0}, I)$ and the Stochastic ML objective (refer to [21]) is used as the ML objective function for DOA estimation. Here MSE is calculated as the sum of mean-squared errors of $\hat{\psi}_1$ and $\hat{\psi}_2$.

Fig. 8 compares the performance of the different methods. As in the case of frequency estimation, Method 1 is better than the root-MUSIC. Method 1 is also better than the root-swap method of Shaghaghi and Vorobyov [10] but inferior to ML method. Method 2 is better than Method 1 and achieves threshold SNR of the ML method.

V. CONCLUSION

We proposed two efficient methods for sinusoidal frequency estimation, which can also be used for the DOA problem. A set of at most $5p$ frequencies were chosen around LSP neighborhoods that are derived from the $A(z)$ obtained from root-MUSIC. The search set has at most 5pC_p points. In Method 1, the threshold performance was found to be similar

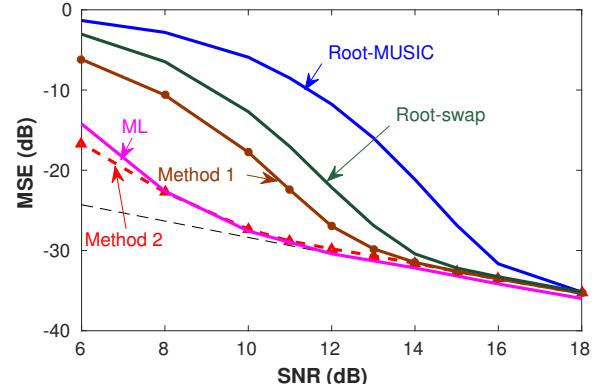


Fig. 8. Estimation performance of different methods for DOA estimation example.

to that of ML for some of the widely used examples. Using the method based on the beamformer function to remove outliers, we proposed Method 2, which was found to either match or improve ML's threshold values for the standard examples (except in one or two cases). Because the evaluation is only over a small set of frequency points, the method is computationally efficient.

APPENDIX

A. Derivation of LSP Properties

LSP properties for the real-valued case is given in [11]. The same approach is applied here for the complex case.

The LSP polynomials $P(z)$ and $Q(z)$ can be rewritten as

$$A(z) \left(1 \pm z^{-(M+1)} \frac{A^*(1/z^*)}{A(z)} \right) = A(z) (1 \pm H(z))$$

where $H(z)$ is an all-pass term. If $A(z)$ is minimum-phase, then $P(z) = 0$ or $Q(z) = 0$ requires $H(z) = \pm 1$, that is, $|H(z)| = 1$. Let $A(z) = \prod_{i=1}^M (1 - z_i z^{-1})$. Then

$$|H(z)| = \frac{1}{|z|} \prod_{i=1}^M \frac{|1 - z_i^* z|}{|z - z_i|} \quad (7)$$

Considering the the k -th term,

$$\begin{aligned} |1 - z_k^* z|^2 - |z - z_k|^2 &= |z_k|^2 |z|^2 + 1 - |z_k|^2 - |z|^2 \\ &= (1 - |z_k|^2)(1 - |z|^2) \end{aligned} \quad (8)$$

Since $|z_k| < 1$ (minimum phase) and if $|z| < 1$, Eq. (8) gives $|1 - z_k^* z|^2 > |z - z_k|^2$. Hence, since $1/|z| > 1$, we get $|H(z)| > 1$. Similarly, for $|z| > 1$, $|H(z)| < 1$. From these two results we see that $|H(z)| = 1$ only when $|z| = 1$.

- Thus all the zeros of $P(z)$ and $Q(z)$ lie on the unit circle.
- Location of these zeros can be found out using the $n\pi$ crossings of $\angle H(e^{j\omega})$.

Let $z_i = r_i e^{j\omega_i}$ represent the roots of $A(z)$, then

$$\angle H(e^{j\omega}) = -(M+1)\omega - \sum_{i=1}^M 2 \tan^{-1} \frac{r_i \sin(\omega - \omega_i)}{1 - r_i \cos(\omega - \omega_i)} \quad (9)$$

$\angle H(e^{j\omega})$ is a monotonically decreasing function (its derivative can be proved to be negative for minimum phase $A(z)$) [11]. Thus, the interlacing behavior of roots of $P(z)$ and $Q(z)$ is also evident, since $H(e^{j\omega})$ takes on values +1 and -1 alternately corresponding to the $n\pi$ crossings of $\angle H(e^{j\omega})$.

B. Derivation of Eq. (3)

Let $A(z) = 1 + \sum_{k=1}^M a_k z^{-k}$, then

$$P(e^{j\omega}) = 1 + \left(\sum_{k=1}^M (a_k + a_{M+1-k}^*) e^{-j\omega k} \right) + e^{-j\omega(M+1)}$$

Taking its conjugate,

$$\begin{aligned} P^*(e^{j\omega}) &= 1 + \left(\sum_{k=1}^M (a_k^* + a_{M+1-k}) e^{j\omega k} \right) + e^{j\omega(M+1)} \\ &= e^{j\omega(M+1)} P(e^{j\omega}) \end{aligned}$$

Similarly, it is easy to show that

$$Q^*(e^{j\omega}) = -e^{j\omega(M+1)} Q(e^{j\omega})$$

Hence,

$$P(e^{j\omega}) Q^*(e^{j\omega}) + P^*(e^{j\omega}) Q(e^{j\omega}) = 0 \quad (10)$$

We know that $A(z) = (P(z) + Q(z))/2$. Therefore,

$$|A(e^{j\omega})|^2 = [P(e^{j\omega}) + Q(e^{j\omega})] \cdot [P^*(e^{j\omega}) + Q^*(e^{j\omega})] \div 4 \quad (11)$$

From Eq. (10) and Eq. (11), we get

$$|A(e^{j\omega})|^2 = (|P(e^{j\omega})|^2 + |Q(e^{j\omega})|^2) \div 4 \quad (12)$$

We know that the roots of $P(z)$ and $Q(z)$ lie on the unit circle. Let $e^{j\omega_{P_1}}, e^{j\omega_{P_2}}, \dots, e^{j\omega_{P_{M+1}}}$ be the roots of $P(z)$ and $e^{j\omega_{Q_1}}, e^{j\omega_{Q_2}}, \dots, e^{j\omega_{Q_{M+1}}}$ be the roots of $Q(z)$.

Then

$$\begin{aligned} P(z) &= \prod_{k=1}^{M+1} (1 - e^{j\omega_{P_k}} z^{-1}) \\ |P(e^{j\omega})|^2 &= \prod_{k=1}^{M+1} |1 - e^{-j(\omega - \omega_{P_k})}|^2 \\ &= 4^{M+1} \prod_{k=1}^{M+1} \sin^2 \left(\frac{\omega - \omega_{P_k}}{2} \right) \end{aligned} \quad (13)$$

Similarly we get,

$$|Q(e^{j\omega})|^2 = 4^{M+1} \prod_{k=1}^{M+1} \sin^2 \left(\frac{\omega - \omega_{Q_k}}{2} \right) \quad (14)$$

Substituting Eq. (13) and Eq. (14) in Eq. (12), we get the final result,

$$|A(e^{j\omega})|^2 = \frac{\prod_{k=1}^{M+1} \sin^2 \left(\frac{\omega - \omega_{P_k}}{2} \right) + \prod_{l=1}^{M+1} \sin^2 \left(\frac{\omega - \omega_{Q_l}}{2} \right)}{4^{-M}} \quad (15)$$

REFERENCES

- [1] S. Kay and J. Marple, S.L., "Spectrum analysis - a modern perspective," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1380–1419, Nov 1981.
- [2] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Englewood Cliffs: Prentice Hall, 1987.
- [3] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.
- [4] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *Signal Processing Magazine, IEEE*, vol. 13, no. 4, pp. 67–94, Jul 1996.
- [5] Q. Cheng and Y. Hua, "Detection of cisoids using least square error function," *IEEE Transactions on Signal Processing*, vol. 45, no. 6, pp. 1584–1590, Jun 1997.
- [6] D. Rife and R. Boorstyn, "Multiple tone parameter estimation from discrete-time observations," *Bell System Technical Journal*, vol. 55, no. 9, pp. 1389–1410, Nov 1976.
- [7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, Mar 1986.
- [8] A. J. Barabell, "Improving the resolution performance of eigenstructure-based direction-finding algorithms," in *Intl. Conf. on Acoust., Speech, and Sig. Process., ICASSP'83*, vol. 8. IEEE, 1983, pp. 336–339.
- [9] S. Umesh, "Fast maximum likelihood estimation of parameters in crowded signal environments," Ph.D. dissertation, Univ. of Rhode Island, 1993.
- [10] M. Shaghaghi and S. A. Vorobyov, "Subspace leakage analysis and improved DOA estimation with small sample size," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3251–3265, June 2015.
- [11] F. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Intl. Conf. on Acoust., Speech, and Sig. Process., ICASSP'84*, vol. 9, Mar 1984, pp. 37–40.
- [12] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.
- [13] D. W. Tufts and R. Kumaresan, "Estimation of frequencies of multiple sinusoids: Making linear prediction perform like maximum likelihood," *Proceedings of the IEEE*, vol. 70, no. 9, pp. 975–989, 1982.
- [14] A. B. Gershman, J. Ringelstein, and J. F. Böhme, "Removing the outliers in root-MUSIC via conventional beamformer," *Signal Processing*, vol. 60, no. 2, pp. 251 – 254, 1997. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0165168497800102>
- [15] V. Vasylyshyn, "Removing the outliers in root-MUSIC via pseudo-noise resampling and conventional beamformer," *Signal Processing*, vol. 93, no. 12, pp. 3423 – 3429, 2013, special Issue on Advances in Sensor Array Processing in Memory of Alex B. Gershman. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S016516841300217X>
- [16] C. Qian, L. Huang, and H. C. So, "Improved unitary root-MUSIC for DOA estimation based on pseudo-noise resampling," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 140–144, Feb 2014.
- [17] I. V. McLoughlin, "Line spectral pairs," *Signal Processing*, vol. 88, no. 3, pp. 448 – 467, 2008. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0165168407003167>
- [18] A. Lepschy, G. Mian, and U. Viaro, "A note on line spectral frequencies [speech coding]," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 8, pp. 1355–1357, Aug 1988.
- [19] Tom Bäckström and Carlo Magi, "Properties of Line Spectrum Pair Polynomials—A Review," *Signal Processing*, vol. 86, no. 11, pp. 3286–3298, 2006.
- [20] N. Sugamura and N. Farvardin, "Quantizer design in LSP speech analysis-synthesis," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 432–440, Feb 1988.
- [21] P. Stoica and A. Nehorai, "Performance study of conditional and unconditional direction-of-arrival estimation," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 38, no. 10, pp. 1783–1795, Oct 1990.

Adaptive Multiple-pixel Wide Seam Carving

Diptiben Patel

Electrical Engineering

Indian Institute of Technology Gandhinagar
Gandhinagar, India 382355

Email: diptiben.patel@iitgn.ac.in

Srivathsan Shanmuganathan

Computer Engineering

University of Jaffna
Jaffna, Sri Lanka

Email: srivathsan.s@iitgn.ac.in

Shanmuganathan Raman

Electrical Engineering

Indian Institute of Technology Gandhinagar
Gandhinagar, India 382355

Email: shanmuga@iitgn.ac.in

Abstract—Content-aware image retargeting methods address the resizing of an image to be displayed on devices having different aspect ratios and resolutions. Seam carving method is an effective image retargeting method which suffers from high computational complexity. It requires one to find one-pixel wide minimum energy path in either vertical or horizontal direction, called *seam*, to reduce the image size by one pixel. In this paper, we propose an acceleration of the seam carving method by expanding the width of the seam making it multiple-pixel wide seam carving. The two types of energies: one corresponding to the pixels to be removed and another corresponding to the pixels across the multiple-pixel wide seam, increase as the width of the seam increases. In order to prevent the increase in these energies, we make the width of the seam adaptive as a function of the number of iterations. We find the width of a seam for each iteration as a prior for the seam carving process using a set of maximum energy seams in an orthogonal direction to the seam carving process. Qualitative and quantitative results prove that the proposed method performs faster and better than the other state-of-the-art image retargeting operators.

I. INTRODUCTION

With the advancement of technology, internet has become a comprehensive tool used for various purposes. A great deal of information on the internet today is presented through images. Access of these images using different screen sizes and display devices such as laptops, desktop PCs, cell phones, and tablets has led to the requirement of resizing the images. The standard resizing operators involve scaling or cropping the image to fit the screen size. However, scaling can be applied only uniformly, and it is not aware of the salient content of our interest. Cropping is the process consisting of removal of unwanted areas. Since these methods only consider the geometric constraints without considering the content of the image, we need an effective content-aware image resizing operator to achieve better results.

Seam carving is one of the many methods for content-aware image retargeting [1]. In order to change the size of an image, the lowest energy path of one-pixel width, called seam, is found from top to bottom or left to right of an image. The lowest energy seam is removed or inserted to resize the image to make it fit the screen size. However, the computational complexity of this method is high. For each iteration, we need to update the energy map of the image and find the lowest energy seam using dynamic programming.

978-1-5386-9286-8/19/\$31.00 © 2019 IEEE

This paper aims to speed up the seam carving process by finding multiple connected seams and evaluating the performance by removing them instead of the single-pixel wide seam removal process. While removing or inserting a seam of single-pixel width passing through low energy region, there is a very less probability of creating an edge of higher magnitude. The reason for the low probability is that the neighboring pixels of low energy region in an image tend to have similar features. However, when we remove a consecutive array of pixels from a row/column, spatially distant pixels in a row/column become neighbors and they may create higher magnitude edge which is undesirable information and not consistent with the original image. We minimize the magnitude of this undesirable information by using the notion of forward energy. The forward energy takes care of the minimum energy to be inserted along with the minimum energy to be removed [2]. The major contributions of this paper are as follows.

- 1) We propose a seam carving technique that accelerates the retargeting process with an increase in the width of the seams while achieving good image retargeting performance comparable to the other state-of-the-art image retargeting operators.
- 2) The energy contained in a seam increases with an increase in the removal of multiple-pixel wide seams. The increased energy leads to distortion across the width of the seam. To prevent the image distortion, we adaptively vary the width of the seam according to the image content.
- 3) In order to vary the width of the seam, energy variation across the number of iterations needs to be processed. The processing of the energy vector is computationally time-consuming. We propose a clustering-based approach on the set of maximum energy seams to decide the width of a seam in each iteration as a prior for the seam carving process.

The rest of the paper is structured as follows. In Section II, we present a brief overview of the image retargeting works. Section III explains about the methodology we propose for the retargeting process. Results are discussed in Section IV. We conclude the paper in section V with the future scope.

II. RELATED WORK

Many works have been developed to perform cropping and scaling while being aware of the image contents. Content-

aware cropping methods find cropping window of the target display size which preserves the salient region and removes a distracting region, if any, available in the image. The best cropping window is found using saliency [3], gaze detection [4], aesthetic features [5], [6], or view of expert photographers [7]. Content-aware scaling methods find a local scaling factor by preserving the aspect ratio of the salient content [8], [9].

Content-aware image retargeting methods are classified as continuous and discrete retargeting methods [10]. Continuous image retargeting methods find the mapping of a quadrangle or triangular grid vertices from the source image area to the target image area allowing less deformation in the salient or important region of an image and more distortion in non-salient or background region of an image [11]–[16]. Discrete image retargeting methods find a set of low energy pixels which can be removed or added to resize the image. The low energy pixels are defined using different image features such as shift map [17], depth information [18], 3D saliency [19], and object occlusions [20]. As no single operator is able to perform well on all the images, integration and optimization of multiple operators such as scaling, cropping, seam carving, and warping are exploited for retargeting the image in [21]–[26].

Seam carving finds an 8-connected low energy seam in either vertical or horizontal direction [1]. The input image is resized by one column (row) by removing or adding the vertical (horizontal) seam while preserving the total energy of an image. The seam carving based image retargeting techniques are compared in [27]. As we accelerate the seam carving method using multiple-pixel wide seam carving with adaptive seam width as a prior in the proposed work, we briefly discuss the variations of the seam carving methods closely relevant to the acceleration aspect. Han *et al.* have detected multiple disconnected seams using graph cut formulation on voxels [28]. The distance between two adjacent seams makes all seams to spread uniformly over the image. This uniformity of seams is in contrast to the observation of seam carving that more seams tend to pass through the low energy region of an image. Seam importance, region smoothness, and seam shape prior information, are used to design an energy function for image retargeting purpose in [29]. In [30], multiple seams are inserted using inpainting around a single pixel wide minimum energy seam. The inpainting process does not assure the minimum energy criteria for the retargeting process. Blocks are considered as a seam in order to remove or add multiple connected pixels in [31]. Creation of new edges and fixed block size are the limitations of this work. Conger *et al.* have combined seam carving and wavelet transform in order to remove the seam of multiple-pixel width around single pixel wide minimum energy seam [32].

III. PROPOSED MULTIPLE-PIXEL WIDE SEAM CARVING

The primary goal of the proposed approach is to accelerate the seam carving process by removing multiple-pixel connected seams in such a way that the quality of the retargeted image is maintained as that of the naïve seam carving

approach. It is observed from the seam carving process in [1] that the consecutive seams tend to pass from the same low energy region until there is no possibility to expand further. Most of the multiple seam carving methods either spread the seam uniformly over the image or suffer from the limitations of fixed seam width over all the iterations. We propose multiple-pixel wide seam carving method where the seam width for each iteration is calculated as a prior to the seam carving process. The input to the proposed approach is an image \mathbf{I} of size $m \times n$, desired size of retargeted image $m \times n'$, and the initial width of multiple-pixel connected seams K . The output of the proposed approach is a retargeted image of size $m \times n'$. The block diagram of the proposed approach is shown in Fig. 1. Without loss of generality, we change the width n using the vertical seam carving process. However, one can similarly go for horizontal seam carving after rotating the energy map and the input image by 90° and performing the image retargeting.

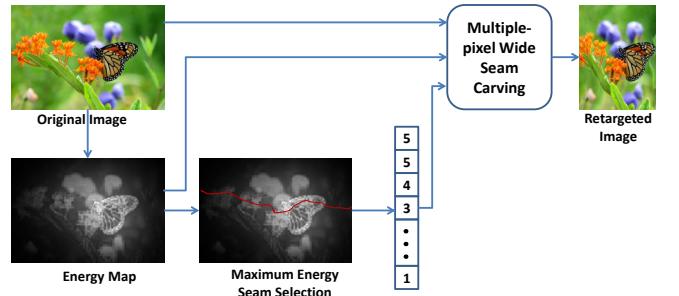


Fig. 1: Proposed multiple-pixel wide seam carving.

A. Energy Map

The energy map \mathbf{E} of an image states the importance of a pixel in an image. There are various features to be considered for an energy map for the image retargeting application like gradient, saliency, object map, depth map, memorability, to name a few [1], [18], [19], [26], [33]. We have used the saliency detection model mentioned in [34] as the energy map for the proposed approach. The intensity, color and texture features in the Discrete Cosine Transform (DCT) domain are fused to define the saliency of each 8×8 blocks. The Hausdorff distance between these features in one block and those of in all the other blocks derives the energy map of the image.

B. Multiple-pixel Seam Width As a Prior

In order to alter the size of an image \mathbf{I} , we propose to remove k -pixel wide seam of lowest energy in a single iteration. Here, k is the number of consecutive pixels in a row. We find the value of k for each iteration t beforehand to reduce the overhead of tracking the energy in every iteration. The vector \mathbf{k} as a function of iteration t is defined such that $\sum_{t=1}^Z k(t) = n - n'$. Z is the total number of iterations required to achieve the retargeted image of size $m \times n'$.

In order to find the vector \mathbf{k} , we find the maximum energy seam in the opposite direction of the image retargeting process (in the horizontal direction for vertical seam carving and vice-versa). After finding the maximum energy seam, we cluster

the set of consecutive pixels along the maximum energy seam having low energy. We define the cost of a seam as in Eq. (1).

$$e(\mathbf{s}) = e(\mathbf{I}_s) = \sum_{i=1}^n \mathbf{E}(\mathbf{I}(s_i)) \quad (1)$$

Here, \mathbf{E} is the energy map. Then, we look for the seam s_{\max} that maximizes this seam cost as shown in Eq. (2).

$$s_{\max} = \arg \max_{\mathbf{s}} (e(\mathbf{s})) = \arg \max_{\mathbf{s}} \sum_{i=1}^n \mathbf{E}(\mathbf{I}(s_i)) \quad (2)$$

Similar to finding a low energy seam, the maximum energy seam can be found using dynamic programming by traversing the image from the second column to the last column and computing the cumulative maximum energy \mathbf{M} for all the possible connected seams for each (i, j) as given by Eq. (3).

$$\begin{aligned} M(i, j) &= E(i, j) + \\ &\max(M(i-1, j-1), M(i, j-1), M(i+1, j-1)) \end{aligned} \quad (3)$$

At the end of this process, the maximum value in the last column of \mathbf{M} will indicate the end of the horizontal seam of maximum energy. Then we backtrack the maximum energy of \mathbf{M} from the last column to the first column to find the path of the maximum energy seam. To define the width of a seam in each iteration of the vertical seam removal, we can not rely only on a single horizontal seam of maximum energy. We find the number of maximum energy seams (P) and calculate the average energy of them. The average of P maximum energy seams is given by Eq. (4).

$$s_{\text{avg}} = \frac{1}{P} \sum_{p=1}^P s_{\max}^p \quad (4)$$

Once we have found the average energy of the P number of maximum energy seams, we sort the calculated average energy in the ascending order. The vector \hat{s}_{avg} contains the sorted energy values of s_{avg} in ascending order. According to the required size of the output image, we find a threshold value from the vector \hat{s}_{avg} . If the width of the original image of size $m \times n$ needs to be retargeted to $m \times n'$, then the threshold value T is given by Eq. (5).

$$T = \hat{s}_{\text{avg}}(n - n') \quad (5)$$

The assumption behind this approach is that the low energy seams along the vertical direction cross the maximum energy seams at the pixels with low energy.

By calculating the value of T , we estimate the highest energy of the pixel across the maximum energy seam that we can remove during the retargeting process. In order to preserve the energy of retargeted image, we can remove pixels with energy below the threshold value T using a multiple-pixel wide seam carving. After finding the threshold T from \hat{s}_{avg} , we define the vector \mathbf{r} from the vector s_{avg} as defined in Eq. (6).

$$\mathbf{r}(j) = \begin{cases} 1, & \text{if } s_{\text{avg}} \leq T \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

Here, $j = 1, 2, 3, \dots, n$ and $\mathbf{r} \in \{0, 1\}^n$. From the vector \mathbf{r} , we cluster the number of consecutive ones adjacent to each other and form a vector $\hat{\mathbf{k}}$ by arranging the number of consecutive ones in descending order. We clip the value in $\hat{\mathbf{k}}$ to the maximum value K in order to reduce the distortion. The vector \mathbf{k} indicating the width of the seam in every iteration is defined in Eq. (7).

$$\mathbf{k}(t) = \begin{cases} \hat{\mathbf{k}}(t), & \text{if } \hat{\mathbf{k}}(t) \leq K \\ K, & \text{otherwise} \end{cases} \quad (7)$$

Here, $t = 1, 2, \dots, Z$. According to the value mentioned in the vector \mathbf{k} , we perform the multiple-pixel wide seam carving.

C. Multiple-pixel Wide Seam Carving

Using an adaptive width \mathbf{k} of a seam, we remove or insert $\mathbf{k}(t)$ pixel wide seam in a t^{th} iteration in order to alter the size of an image \mathbf{I} by $\mathbf{k}(t)$ columns or rows. Here, \mathbf{k} is the number of lowest energy consecutive pixels in a single row (column) for the vertical (horizontal) seam detection process. After the removal of array pixels during the seam carving process, pixels at a spatial distance of \mathbf{k} -pixel become new neighbors as shown in Fig. 2. These pixels may relate to distant features creating new edges of significant magnitude. These newly constructed edges insert energy, named as forward energy, to an image. This forward energy needs to be minimized in order to preserve the total energy of an image. We measure the cost of the forward energy by taking absolute forward differences between the pixels which are going to become new neighbors after the seam removal or insertion.

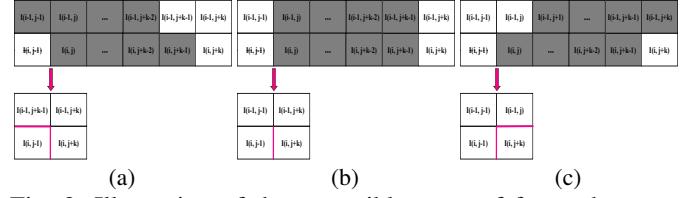


Fig. 2: Illustration of three possible costs of forward energy for vertical multiple-pixel wide seam. After the removal of a seam, new edges created are shown with pink color.

The minimum cumulative energy for the vertical seam is calculated by traversing the image starting from the first row to the last row for all possible minimal energy seams of \mathbf{k} -pixel width. We find the vertical \mathbf{k} -pixel wide seam of minimum energy by backtracking in the upper direction from the minimum value of cumulative energy. We repeat the seam removal or insertion process until the image of the target display size is achieved. The adaptive multiple-pixel wide seam carving process is summarized in Algorithm 1.

IV. RESULTS AND DISCUSSION

A. Dataset

We have used RetargetMe dataset [10] for evaluating the performance of the proposed approach. It contains a total of 80 images having attributes such as people, faces, lines, clear edges, evident foreground objects, texture elements, specific

geometric structures, and symmetric objects. We compared the results of our approach for change in aspect ratio of 0.75 and 0.50 with the other state-of-the-art image retargeting operators such as Cropping (CR), Streaming Video (SV) [35], Multi-Operator (MULTIOP) [22], Seam Carving (SC) [2], Scaling (SCL), Shift-Maps (SM) [17], Scale-and-Stretch (SNS) [8], and Nonhomogeneous Warping (WARP) [36] from different classifications like traditional, continuous, discrete, and combined image retargeting operators.

B. Quality Assessment Measure

We have evaluated the performance of the proposed method using a multiple-level feature (MLF) based quality measure proposed in [37]. The quality of a retargeted image is defined using a low-level feature: aspect ratio similarity (Q_{ARS}), mid-level feature: edge group similarity (Q_{EGS}) and high-level feature: face block similarity (Q_{FBS}). Here, $\{Q_{ARS}, Q_{EGS}, Q_{FBS}\} \in [0, 1]$. The multiple-level features are complementary as they quantify different aspects of quality degradation in the retargeted image. The three features are fused using a support vector machine (SVM) regression model trained on the subjective data of 37 images provided with the RetargetMe dataset [10]. The trained SVM model outputs the ranking order of an image retargeted with different operators including the proposed method.

C. Discussion

We have compared the performance of the proposed method with the other state-of-the-art image retargeting operators in both quantitative and qualitative manner. The number of maximum energy seam in the orthogonal direction is $P = 5$. We evaluated the proposed image retargeting method for initial seam width values $K = 3, 5$, and 7 . TABLE I and TABLE II list the Q_{ARS} , Q_{EGS} , and Q_{FBS} features for some of the retargeted images for change in aspect ratio of 0.75 and 0.50, respectively. The highest and the second highest values of respective features for an image are shown in red and blue colors, respectively. We can observe from TABLE I and TABLE II that the proposed method performs better than the other state-of-the-art image retargeting operators for $K = 3$. Also, increasing the initial seam width K compromises little in terms of different features as seen in TABLE I and TABLE II. Q_{FBS} is a face block similarity feature defined for the images having a face or 1 otherwise. We only report Q_{FBS} feature values for the images containing a face(s) such as *Woman* and *Girls*.

We have analyzed the relative ranking of the different image retargeting operators including the proposed method. Fig. 3(a) shows the number of images (in percentage) ranked for the proposed method with different initial seam width $K = 3, 5$, and 7 . Observing Fig. 3(a), 91%, 61% and 48% of images from the total images are ranked top 3 for initial seam width $K = 3, 5$, and 7 , respectively. Fig. 3(b) shows the number of images (in percentage) ranked top 1 for the different image retargeting operators. One can observe that the proposed method outperformed the other image retargeting

Algorithm 1: Image retargeting using adaptive multiple-pixel wide seams.

Input : Original image \mathbf{I} of size $m \times n$, Energy map E , Retargeted image size $m \times n'$, Number of maximum energy seams P , Initial seam width K .

Output: Retargeted image \mathbf{I}_R .

- 1 $K \leftarrow K$, $P \leftarrow P$, $\mathbf{I}_R \leftarrow \mathbf{I}$.
- 2 $s_{avg} = \frac{1}{P} \sum_{p=1}^P s_{max}^p$.
- 3 $\hat{s}_{avg} = s_{avg}$ in ascending order.
- 4 $T = \hat{s}_{avg}(n - n')$.
- 5 **for** $j \leftarrow 1$ **to** n **do**
- 6 $r(j) = \begin{cases} 1, & \text{if } s_{avg} \leq T. \\ 0, & \text{otherwise.} \end{cases}$
- 7 **end**
- 8 $\hat{k} = \text{number of consecutive ones in } r \text{ in descending order such that, } \sum_{t=1}^Z k(t) = n - n'$.
- 9 **for** $t \leftarrow 1$ **to** Z **do**
- 10 $k(t) = \begin{cases} \hat{k}(t), & \text{if } \hat{k}(t) \leq K. \\ K, & \text{otherwise.} \end{cases}$
- 11 **end**
- 12 **foreach** k **do**
- 13 Remove or insert k -pixel wide seam from \mathbf{I}_R .
- 14 **end**

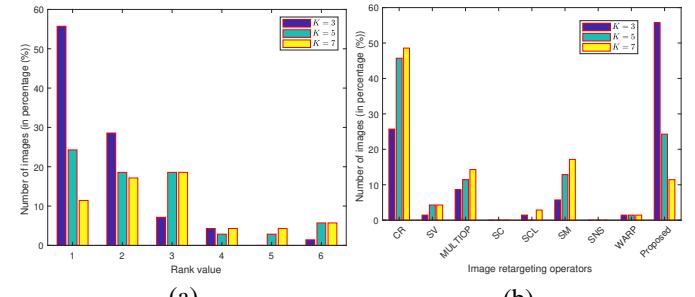


Fig. 3: Percentage number of images ranked for proposed method against the other state-of-the-art image retargeting operators.

operators for 56% of the total images for seam width $K = 3$. As the width of seam K increases, the seams of higher width passing through the image distort edges leading to reduction in edge group similarity (Q_{EGS}) feature which is also evident from TABLE I and TABLE II. Also, edge group similarity (Q_{EGS}) feature is always best preserved for cropping (CR) operator because of no local distortion of edges rather than the loss of content from the border.

The visual quality of the retargeted images for different image retargeting operators is shown in Fig. 4 and Fig. 5. Fig. 4 shows the retargeted images for change in aspect ratio of 0.75 and Fig. 5 shows the retargeted images for change in aspect ratio of 0.50. The proposed method is able to preserve the geometric structure and salient objects better than the other

TABLE I: Comparison of the proposed method with other state-of-the-art image retargeting operators using Multiple-level features (MLF) for change in aspect ratio of 0.75.

Image Name	Feature	CR	SV [35]	MULTIOP [22]	SC [2]	SCL	SM [17]	SNS [8]	WARP [36]	Proposed ($K = 3$)	Proposed ($K = 5$)	Proposed ($K = 7$)
Battleship	Q_{ARS}	0.9421	0.9568	0.9559	0.8992	0.9526	0.9662	0.9503	0.9388	0.9582	0.9539	0.9332
	Q_{EGS}	0.8970	0.8226	0.8343	0.8304	0.8281	0.8620	0.8206	0.8343	0.9006	0.7937	0.8215
Manga	Q_{ARS}	0.8737	0.9451	0.9528	0.9340	0.9450	0.9212	0.9329	0.9292	0.9570	0.9447	0.9551
	Q_{EGS}	0.9303	0.8410	0.8675	0.8724	0.8399	0.8823	0.8553	0.8650	0.8859	0.8901	0.8813
Woman	Q_{ARS}	0.9394	0.9502	0.9292	0.8810	0.9409	0.8729	0.9564	0.9453	0.9563	0.9558	0.9409
	Q_{EGS}	0.9047	0.8522	0.8630	0.8914	0.8356	0.8590	0.8580	0.8778	0.9116	0.9073	0.8971
	Q_{FBS}	1	0.9900	0.9882	1	0.9571	0.9910	0.9983	0.9969	1	0.9986	1
Girls	Q_{ARS}	0.9502	0.9510	0.9519	0.9164	0.9473	0.9204	0.9392	0.9405	0.9573	0.9560	0.9263
	Q_{EGS}	0.8929	0.8417	0.8549	0.8504	0.8384	0.8857	0.8386	0.8449	0.8846	0.8768	0.8722
	Q_{FBS}	1	0.9884	0.9800	0.9776	0.9550	0.9934	0.9625	0.9554	0.9943	0.9930	0.9982

TABLE II: Comparison of the proposed method with other state-of-the-art image retargeting operators using Multiple-level features (MLF) for change in aspect ratio of 0.50.

Image Name	Feature	CR	SV [35]	MULTIOP [22]	SC [2]	SCL	SM [17]	SNS [8]	WARP [36]	Proposed ($K = 3$)	Proposed ($K = 5$)	Proposed ($K = 7$)
Family	Q_{ARS}	0.8659	0.8393	0.8327	0.7840	0.7663	0.7840	0.8363	0.8231	0.8855	0.8654	0.8072
	Q_{EGS}	0.8969	0.7967	0.8270	0.7838	0.7828	0.8877	0.7905	0.8069	0.9215	0.8395	0.8660
Eagle	Q_{ARS}	0.8758	0.8374	0.8399	0.8205	0.7845	0.8739	0.8417	0.8774	0.8958	0.8906	0.8841
	Q_{EGS}	0.9010	0.7409	0.7543	0.7557	0.7269	0.8171	0.7552	0.7986	0.8774	0.8593	0.8978
Fishing	Q_{ARS}	0.8040	0.8158	0.7844	0.7749	0.7663	0.7986	0.7695	0.7582	0.8248	0.8242	0.8187
	Q_{EGS}	0.9234	0.7582	0.7736	0.7732	0.7465	0.8489	0.7305	0.7687	0.8637	0.8763	0.8715
Orchid	Q_{ARS}	0.7608	0.8328	0.8335	0.8113	0.7731	0.5201	0.8132	0.8062	0.8870	0.8815	0.8228
	Q_{EGS}	0.8939	0.7772	0.7851	0.7790	0.7651	0.8437	0.7698	0.7876	0.8660	0.8469	0.8160
Waterfall	Q_{ARS}	0.8091	0.8166	0.7965	0.8076	0.7757	0.8438	0.7880	0.7905	0.8597	0.8491	0.8089
	Q_{EGS}	0.8963	0.7274	0.7699	0.7838	0.7438	0.8199	0.7353	0.78190	0.8639	0.8794	0.8859

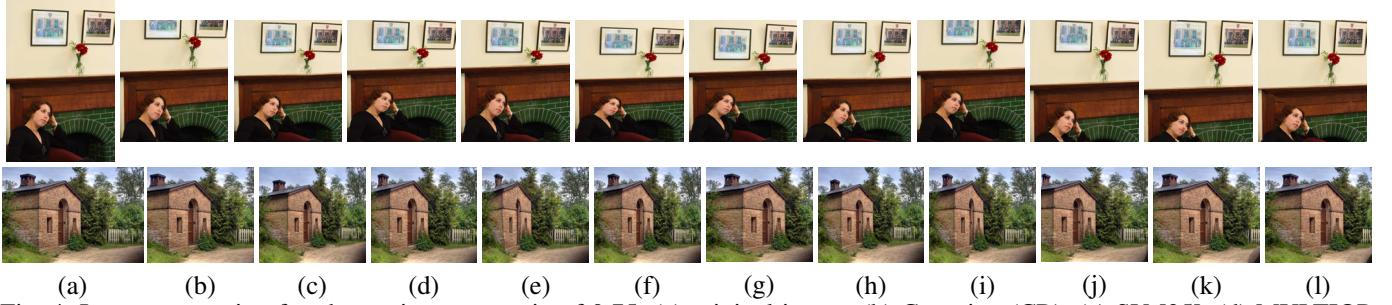


Fig. 4: Image retargeting for change in aspect ratio of 0.75: (a) original image, (b) Cropping (CR), (c) SV [35], (d) MULTIOP [22], (e) SC [2], (f) Scaling (SCL), (g) SM [17], (h) SNS [8], (i) Warp [36], (j) Proposed ($K = 3$), (k) Proposed ($K = 5$), and (l) Proposed ($K = 7$).

state-of-the-art image retargeting operators. The quantitative values for the cropping operator are comparable to those of the proposed method, but the proposed method is able to preserve salient objects more efficiently as evident from Fig. 4 and Fig. 5 (See photo frame in *Woman*, background mountains in *Deck*, orange colored flowers in *Butterfly* image). For different initial seam width K , we calculated the time required in seconds for retargeting an image by one column/row. TABLE III shows the mean and standard deviation of the time required for the change in the size of an image by a single column/row. We consider the time per single column/row because the dataset contains images of various sizes. As the seam width K increases, the time required for the retargeting process reduces considerably. Also, a decrease in the standard deviation shows more reliability in the average time required.

TABLE III: Time Analysis of the proposed method (in sec).

Parameter	$K = 1$	$K = 3$	$K = 5$	$K = 7$
Mean	4.26	1.48	0.78	0.60
Standard deviation	2.44	0.91	0.39	0.29

V. CONCLUSION

We have accelerated the seam carving process, an important subclass of the discrete image retargeting techniques. We have retargeted the image by multiple-pixels in an iteration rather than a single pixel, using multiple-pixel wide seam carving. The width of the seam is a crucial factor for preserving the total energy of the image. We have proposed a method to find an adaptive multiple-pixel seam width as a function of the iteration number as a prior to the seam carving process. We have retargeted the images using a width of a seam calculated apriori for every iteration. We have compared the proposed image retargeting method with the other state-of-the-art image retargeting operators and have shown that the proposed method performs better than the other image retargeting operators both qualitatively and quantitatively. Also, we have shown that by increasing the initial seam width, the time required to retarget an image to the target display size reduces drastically. We would like to address the constraint of connected multiple-pixel seams as future work.



Fig. 5: Image retargeting for change in aspect ratio of 0.50: (a) original image, (b) Cropping (CR), (c) SV [35], (d) MULTIOP [22], (e) SC [2], (f) Scaling (SCL), (g) SM [17], (h) SNS [8], (i) Warp [36], (j) Proposed ($K = 3$), (k) Proposed ($K = 5$), and (l) Proposed ($K = 7$).

REFERENCES

- [1] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” in *ACM Trans. on graphics*, 2007, vol. 26, p. 10.
- [2] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Trans. on graphics*, vol. 27, no. 3, pp. 16, 2008.
- [3] B. Suh, H. Ling, B.B. Bederson, and D.W. Jacobs, “Automatic thumbnail cropping and its effectiveness,” in *Proc. of the ACM symp. on User interface software and technology*, 2003, pp. 95–104.
- [4] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, “Gaze-based interaction for semi-automatic photo cropping,” in *Proc. of the SIGCHI conf. on Human Factors in computing systems*, 2006, pp. 771–780.
- [5] Y. Kao, R. He, and K. Huang, “Automatic image cropping with aesthetic map and gradient energy map,” in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2017, pp. 1982–1986.
- [6] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, “Probabilistic graphlet transfer for photo cropping,” *IEEE Trans. on Image Processing*, vol. 22, no. 2, pp. 802–815, 2013.
- [7] J. Yan, S. Lin, S. Bing Kang, and X. Tang, “Learning the change for automatic image cropping,” in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2013, pp. 971–978.
- [8] Y.S. Wang, C.L. Tai, O. Sorkine, and T.Y. Lee, “Optimized scale-and-stretch for image resizing,” in *ACM Trans. on Graphics*, 2008, vol. 27, p. 118.
- [9] Y. Jin, L. Liu, and Q. Wu, “Nonhomogeneous scaling optimization for realtime image resizing,” *The Visual Computer*, vol. 26, no. 6-8, pp. 769–778, 2010.
- [10] M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir, “A comparative study of image retargeting,” in *ACM trans. on graphics*, 2010, vol. 29, p. 160.
- [11] Y. Guo, F. Liu, J. Shi, Z.H. Zhou, and M. Gleicher, “Image retargeting using mesh parametrization,” *IEEE Trans. on Multimedia*, vol. 11, no. 5, pp. 856–867, 2009.
- [12] G.X. Zhang, M.M. Cheng, S.M. Hu, and R.R. Martin, “A shape-preserving approach to image resizing,” in *Computer Graphics Forum*, 2009, vol. 28, pp. 1897–1906.
- [13] H. Wu, Y.S. Wang, K.C. Feng, T.T. Wong, T.Y. Lee, and P.A. Heng, “Resizing by symmetry-summarization,” *ACM Trans. on Graphics*, vol. 29, no. 6, pp. 159, 2010.
- [14] D. Panozzo, O. Weber, and O. Sorkine, “Robust image retargeting via axis-aligned deformation,” in *Computer Graphics Forum*, 2012, vol. 31, pp. 229–236.
- [15] S.S. Lin, I.C. Yeh, C.H. Lin, and T.Y. Lee, “Patch-based image warping for content-aware retargeting,” *IEEE Trans. on multimedia*, vol. 15, no. 2, pp. 359–368, 2013.
- [16] Y. Kim, S. Jung, C. Jung, and C. Kim, “A structure-aware axis-aligned grid deformation approach for robust image retargeting,” *Multimedia Tools and Appl.*, vol. 77, no. 6, pp. 7717–7739, 2018.
- [17] Y. Pritch, E. Kav-Venaki, and S. Peleg, “Shift-map image editing,” in *IEEE Int. Conf. on Computer Vision*, 2009, pp. 151–158.
- [18] J. Shen, D. Wang, and X. Li, “Depth-aware image seam carving,” *IEEE trans. on cybernetics*, vol. 43, no. 5, pp. 1453–1461, 2013.
- [19] Q. Lu, G. Tao, and Y. Chen, “Image retargeting based on self-learning 3d saliency for content-aware data analysis,” *Multimedia Tools and Appl.*, pp. 1–14, 2017.
- [20] A. Mansfield, P. Gehler, L. Van Gool, and C. Rother, “Scene carving: Scene consistent image retargeting,” in *European Conference on Computer Vision*, 2010, pp. 143–156.
- [21] W. Dong, N. Zhou, J.C. Paul, and X. Zhang, “Optimized image resizing using seam carving and scaling,” in *ACM Trans. on Graphics*, 2009, vol. 28, p. 125.
- [22] M. Rubinstein, A. Shamir, and S. Avidan, “Multi-operator media retargeting,” in *ACM Transactions on graphics*, 2009, vol. 28, p. 23.
- [23] S. Luo, J. Zhang, Q. Zhang, and X. Yuan, “Multi-operator image retargeting with automatic integration of direct and indirect seam carving,” *image and vision computing*, vol. 30, no. 9, pp. 655–667, 2012.
- [24] Y. Fang, Z. Fang, F. Yuan, Y. Yang, S. Yang, and N.N. Xiong, “Optimized multioperator image retargeting based on perceptual similarity measure,” *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 11, pp. 2956–2966, 2017.
- [25] Y. Zhang, Z. Sun, P. Jiang, Y. Huang, and J. Peng, “Hybrid image retargeting using optimized seam carving and scaling,” *Multimedia Tools and Appl.*, vol. 76, no. 6, pp. 8067–8085, 2017.
- [26] W. Dong, N. Zhou, T.Y. Lee, F. Wu, Y. Kong, and X. Zhang, “Summarization-based image resizing by intelligent object carving,” *IEEE trans. on vis. and computer graphics*, vol. 20, no. 1, 2014.
- [27] Z. Yan and H. Chen, “A study of image retargeting based on seam carving,” in *IEEE Int. Conf. on Measuring Tech. and Mechatronics Automation*, 2014, pp. 60–63.
- [28] D. Han, X. Wu, and M. Sonka, “Optimal multiple surfaces searching for video/image resizing-a graph-theoretic approach,” in *IEEE Int. Conf. on Computer Vision*, 2009, pp. 1026–1033.
- [29] D. Han, M. Sonka, J. Bayouth, and X. Wu, “Optimal multiple-seams search for image resizing with smoothness and shape prior,” *The Visual Computer*, vol. 26, no. 6–8, pp. 749–759, 2010.
- [30] D. Domingues, A. Alahi, and P. Vandergheynst, “Stream carving: An adaptive seam carving algorithm,” in *IEEE Int. Conf. on Image Processing*, 2010, pp. 901–904.
- [31] K. Mishiba and M. Ikebara, “Block-based seam carving,” in *IEEE Int. Symp. on Access Spaces*, 2011, pp. 111–115.
- [32] D.D. Conger, M. Kumar, and H. Radha, “Multi-seam carving via seamlets,” in *Image Processing: Algorithms and Systems IX*, 2011, vol. 7870, p. 78700H.
- [33] D. Patel and S. Raman, “Saliency and memorability driven retargeting,” in *IEEE Int. Conf. on Signal Processing and Communications*, 2016, pp. 1–5.
- [34] Y. Fang, Z. Chen, W. Lin, and C.W. Lin, “Saliency detection in the compressed domain for adaptive image retargeting,” *IEEE Trans. on Image Processing*, vol. 21, no. 9, pp. 3888–3901, 2012.
- [35] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, “A system for retargeting of streaming video,” in *ACM Trans. on Graphics*, 2009, vol. 28, p. 126.
- [36] L. Wolf, M. Guttmann, and D. Cohen-Or, “Non-homogeneous content-driven video-retargeting,” *Int. Conf. on Computer Vision*, 2007.
- [37] Y. Zhang, W. Lin, Q. Li, W. Cheng, and X. Zhang, “Multiple-level feature-based measure for retargeted image quality,” *IEEE Trans. on Image Processing*, vol. 27, no. 1, pp. 451–463, 2018.

Design of Discrete Frequency-Coding Waveforms Using Phase-Coded Linear Chirp for Multiuser and MIMO Radar Systems

Arijit Roy*, Debasish Deb†, Harshal B. Nemade* and Ratnajit Bhattacharjee*

*Department of Electronics and Electrical Engineering

Indian Institute of Technology Guwahati, Guwahati, India

{arijit.eee, harshal, ratnajit}@iitg.ac.in

†Defence Research and Development Organisation, Bangalore, India

debudeb@gmail.com

Abstract—To perform successful detection of a target, the waveforms operating in a multiuser radar system and multiple-input-multiple-output (MIMO) radar system should maintain minimum cross-correlations among themselves. In this paper, design of discrete frequency-coding waveform (DFCW) using linear chirp (LC) pulse and pseudo-noise (PN) code sequence is proposed to improve the autocorrelation sidelobe peak (ASP) and cross-correlation peak (CCP) levels of the waveforms. To achieve minimum cross-correlation levels between different waveforms, for a multiuser and MIMO radar system, a two-stage optimization process is proposed. In the first stage, optimization of the frequency-firing sequence of the waveforms is performed and in the second stage, PN-code sequence optimization is performed in order to obtain the overall optimized waveforms. The performance of the proposed waveforms in a multiuser and MIMO radar system is analyzed. In comparison with the performance of multiuser and MIMO radar system designed using other available DFCW, the performances of the radar systems using proposed waveforms are better in terms of maintaining lower ASP and CCP levels. In addition, performance of the proposed waveforms under Doppler frequency variation is also presented.

I. INTRODUCTION

Design of radar waveforms, used for a multiuser radar or multiple-input-multiple-output (MIMO) radar applications, having low autocorrelation sidelobe peaks (ASP) and cross-correlation peaks (CCP) is a challenging task. For this purpose, two widely used techniques to design radar waveforms are phase coding technique and frequency diversity technique. Use of pseudo-noise (PN) code sequences for designing radar signals is presented in [1]. In [2], a combination of linear chirp (LC) and PN-code sequence is presented to improve the cross-correlation performance. The waveform is constructed by repeating the same chirp pulse for N times and the starting polarity of each subpulse is decided according to the values (± 1) of the PN-code sequence. Design of discrete frequency-coding waveform (DFCW), by using subpulses of different frequencies through frequency hopping technique, is presented in [3], [4], [5], [6], [7], [8]. Design of DFCW with fixed frequency pulses (DFCW-FF) for radar applications and its

performance is presented in [3], [4], [5], [6]. The combination of DFCW-FF with phase coding, presented in [5], provides better sidelobe property compared to traditional DFCW-FF. A DFCW-FF with phase randomization of the subpulses is presented in [7]. It is shown that use of random phase for each pulses [7] instead of zero or fixed phase [3] improves ASP and CCP levels and provides better performance compared to the waveforms designed using polyphase sequences [9]. Modification of DFCW by replacing fixed-frequency pulses with linear chirp pulses (DFCW-LC) is presented in [6], [8]. Optimization of the frequency-firing sequence [8] allows improvement in ASP and CCP levels over DFCW-FF technique [3], [4], [6]. Although, design of very limited number of waveforms and associated performance is presented in [3], [4], [5], [8], which is insufficient for many multiuser and MIMO radar applications. Further, the performance of the desired waveforms in presence of all other waveforms i.e. performance of the composite signal in a multiuser radar scenario is not addressed in [2], [3], [4], [5], [8].

To address these shortcomings, in this paper, we have proposed the design of a discrete frequency-coding waveform in which the subpulses are designed as linear chirp pulses and the phase of each pulse is varied according to the values of a PN-code sequence. Further, in order to keep ASP and CCP levels of the proposed waveform lower than the reported designs [3], [5], [7], [8], selection of frequency-firing sequence and PN-code sequence need to be done appropriately which leads to the formulation of a frequency and code sequence optimization problem. As a solution, a two-stage optimization process is proposed to minimize ASP and CCP levels of the composite signals and maximize the contribution of the desired signals. The rest of the paper is organized as follows. Section II presents the design process of the proposed discrete frequency-coding waveform with phase-coded linear chirp pulses (DFCW-PNLC) and the formulation of auto-ambiguity function and cross-ambiguity function to analyze the correlation characteristics of such waveforms. Section III presents an optimization process to minimize ASP and CCP levels of the proposed waveforms when operating in a multiuser and

MIMO radar environment. Performance of the waveforms for multiuser and MIMO radar systems is analyzed in Section IV. The results are compared with the performance of the radar systems designed using available waveforms given in [2], [3], [4], [5], [7], [8]. Finally, conclusions are drawn in Section V.

II. DESIGN OF THE PROPOSED DISCRETE FREQUENCY-CODING WAVEFORM USING PHASE-CODED LINEAR CHIRP PULSES (DFCW-PNLC)

A linear chirp signal having bandwidth B and signal duration T can be expressed as [10]

$$g(t) = e^{j2\pi(f_0 t + \frac{1}{2}\mu t^2)}, \quad 0 \leq t \leq T \quad (1)$$

where f_0 is the starting frequency and $\mu = \frac{B}{T}$ is the chirp rate of the signal. Considering a DFCW-LC waveform consists of N chirp subpulses, a total of N waveforms can be formed through frequency-hopping of the available pulses. Thus, n th subpulse of k th waveform can be expressed as

$$g_k(t - (n-1)T) = e^{j2\pi((f_0 + H_k(n)\Delta f - \Delta f)(t - (n-1)T))} \\ \times e^{j\pi\mu(t - (n-1)T)^2}, \quad (n-1)T \leq t \leq nT \quad (2)$$

where $n = 1, 2, \dots, N$, Δf is the frequency step size, H_k is the k th hyperbolic frequency hopping (HFH) sequence [11] for determining the sequence of frequency-firing of chirp pulses. k is the sequence number, n is the index number of the sequence and the corresponding index value is given by $H_k(n)$. The HFH sequence can be formed as [11]

$$H_k(n) = kn^{-1} \pmod{p} \quad (3)$$

where p is a prime number and $1 \leq k, n \leq p-1$. Therefore, a set of $N = p-1$ HFH sequences with each sequence having a length of $N = p-1$ can be generated. Using (2) and (3) k th DFCW-PNLC can be written as

$$s_k(t) = \sum_{i=1}^N c(i) g_k(t - (i-1)T) \quad (4)$$

where $c(i)$ is the PN-code sequence of length N . The instantaneous frequency of one of the DFCW-PNLC is shown in Fig. 1.

For radar applications, we have considered Gold code sequences as the desired PN sequences due to its low cross-correlation property [12]. Gold codes can be generated by combining the outputs of two maximal-length sequence generators [12]. As the length of the PN-code sequence is N , the PN-code set would consist at least N number of Gold code sequences. Further, any DFCW-LC can be encoded with any available PN-code sequences. Thus, using (2) and (4) the general form of k th DFCW-LC encoded with m th PN-code sequence can be expressed as

$$s_k^m(t) = \sum_{i=1}^N c_m(i) g_k(t - (i-1)T) \quad (5)$$

It is considered that no two DFCW-LC, $s_k(t)$ and $s_l(t)$, is encoded with the same PN-code sequence i.e. for two waveforms $s_k^m(t)$ and $s_l^d(t)$, $m \neq d$.

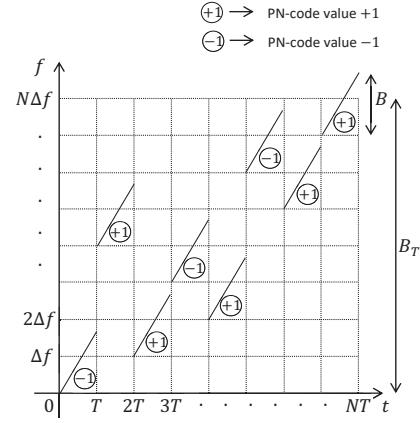


Fig. 1. Instantaneous frequency of one of the DFCW-PNLC.

The performance of the proposed DFCW-PNLC can be evaluated from the analysis of its auto-ambiguity function and cross-ambiguity function [2]. Cross-ambiguity function between two waveforms $s_k^m(t)$ and $s_l^d(t)$ is given by

$$\chi_{s_k^m, s_l^d}(\tau, f_d) = \int_{-\infty}^{\infty} s_k^m(t) s_l^d(t) e^{j2\pi f_d t} dt \quad (6)$$

where $*$ is the complex conjugate operator, f_d is the Doppler frequency shift and τ is the delay time. Using (5), (6) can be simplified and expressed as

$$\begin{aligned} \chi_{s_k^m, s_l^d}(\tau, f_d) &= \int_{-\infty}^{\infty} \sum_{i=1}^N c_m(i) \sum_{i'=1}^N c_d(i') \\ &\times e^{j2\pi((f_0 + H_k(i)\Delta f - \Delta f)(t - (i-1)T) + \frac{1}{2}\mu(t - (i-1)T)^2)} \\ &\times e^{-j2\pi((f_0 + H_l(i')\Delta f - \Delta f)((t - (i'-1)T) - \tau))} \\ &\times e^{-j\pi\mu((t - (i'-1)T) - \tau)^2} e^{j2\pi f_d t} dt \\ &= \sum_{i=1}^N \sum_{i'=1}^N c_m(i) c_d(i') e^{j2\pi f_0[(i'-i)T + \tau]} \\ &\times e^{-j2\pi\Delta f[(i'-i)T + \tau]} \\ &\times e^{j2\pi\Delta f[T(H_l(i')(i'-1) - H_k(i)(i-1)) + H_l(i')\tau]} \\ &\times e^{j\pi\mu[((i-1)^2 - (i'-1)^2)T^2 - 2(i'-1)T\tau - \tau^2]} \\ &\times \int_{-\infty}^{\infty} e^{j2\pi\Delta f[H_k(i) - H_l(i')]t} e^{j2\pi\mu((i'-i)T + \tau)t} dt \\ &\times e^{j2\pi f_d t} dt \\ &= \sum_{i=1}^N \sum_{i'=1}^N c_m(i) c_d(i') e^{j2\pi(f_0 - \Delta f)[(i'-i)T + \tau]} \\ &\times e^{j2\pi\Delta f[T(H_l(i')(i'-1) - H_k(i)(i-1)) + H_l(i')\tau]} \\ &\times e^{j\pi\mu[(i-i')(i+i'-2)T^2 - 2(i'-1)T\tau - \tau^2]} \chi_{ii'} \end{aligned} \quad (7)$$

where $\chi_{ii'} = \int_{t_1}^{t_2} e^{j2\pi(\beta_{ii'}\Delta f + f_d)t} e^{j2\pi\mu((i'-i)T + \tau)t} dt$ and $\beta_{ii'} = H_k(i) - H_l(i')$. In every subpulse duration, for $0 \leq \tau \leq T$, the lower and upper limit of $\chi_{ii'}$ can be given by $t_1 = \tau$ and $t_2 = T$.

For $m = d$, (6) reduces to auto-ambiguity function and using (7) it can be simplified as

$$\chi_{s_k^m, s_k^m}(\tau, f_d) = \int_{-\infty}^{\infty} s_k^m(t) s_k^{m*}(t - \tau) e^{j2\pi f_d t} dt \quad (9)$$

$$= \sum_{i=1}^N \sum_{i'=1}^N c_m(i) c_m(i') e^{j2\pi(f_0 - \Delta f)[(i'-i)T + \tau]} \\ \times e^{j2\pi \Delta f [T(H_k(i')(i'-1) - H_k(i)(i-1)) + H_k(i')\tau]} \\ \times e^{j\pi \mu [(i-i')(i+i'-2)T^2 - \tau^2 - 2(i'-1)T\tau]} \zeta_{ii'} \quad (10)$$

where $\zeta_{ii'} = \int_{t_1}^{t_2} e^{j2\pi(\varphi_{ii'} \Delta f + f_d)t} e^{j2\pi \mu((i'-i)T + \tau)t} dt$ and $\varphi_{ii'} = H_k(i) - H_k(i')$.

III. CROSS-CORRELATION OPTIMIZATION AND WAVEFORM SELECTION FOR MULTIUSER AND MIMO RADAR SYSTEMS

In a multiuser radar system, consider that W users are operating simultaneously where $W < N$. Considering that all the waveforms transmitted from different radar users are different, for a particular user the received signal echo from a target not only consists of the required signal but also have the signals of other radar systems. Thus, minimization of the cross-correlation levels between transmitted signals is required for optimal detection performance. The multiuser radar ambiguity function for v th user can be expressed as

$$\chi_v^{Multiuser}(\tau, f_d) = \int_{-\infty}^{\infty} \sum_{w=1}^W u_w(t) u_v^*(t - \tau) e^{j2\pi f_d t} dt \quad (11)$$

where $u_w(t)$ is the radar signal for w th user. $u_w(t) = s_{k_w}^{m_w}(t)$ and $w = 1, 2, \dots, W$. k_w denotes k th HFH sequence and m_w denotes m th PN-code sequence selected for w th radar signal. Simplifying (11) gives

$$\chi_v^{Multiuser}(\tau, f_d) = \sum_{w=1}^W \int_{-\infty}^{\infty} s_{k_w}^{m_w}(t) s_{k_v}^{m_v*}(t - \tau) e^{j2\pi f_d t} dt \quad (12)$$

The values of k and m of the radar signals corresponding to different users (w) will be different. Thus, replacing k_w , m_w and k_v , m_v , related to user w and v , with the corresponding selected k and m values, the inner integral of (12) takes the form similar to that of (6) and can be evaluated using (8) (An example of k and m values, selected for different values of w is presented in Table II). To achieve optimal detection performance ASP level and CCP level of the correlated output of the composite received radar signal, $\chi_v^{Multiuser}(\tau, f_d)$, should be as low as possible. Similarly, for a MIMO radar system, operates on W radar signals simultaneously through multiple antennas, the received composite signal is processed through a bank of matched filters for processing. And the corresponding MIMO radar ambiguity function can be expressed as [7]

$$\chi_W^{MIMO}(\tau, f_d) = \int_{-\infty}^{\infty} \sum_{w=1}^W \sum_{w'=1}^W u_w(t) u_{w'}^*(t - \tau) e^{j2\pi f_d t} dt \quad (13)$$

$$= \sum_{w=1}^W \sum_{w'=1}^W \int_{-\infty}^{\infty} s_{k_w}^{m_w}(t) s_{k_{w'}}^{m_{w'}*}(t - \tau) e^{j2\pi f_d t} dt \quad (14)$$

Similar to (12), the inner integral of (14) can be solved using (8). Minimum cross-correlation between different transmitted signals allow maximum contribution from each of the signals as achieving complete orthogonality between different received signals is nearly impossible in practice [7]. Thus, appropriate selection of proposed discrete frequency-coding waveforms with phase-coded linear chirp pulses is necessary.

From (8) and (10), for a given B , Δf and T , it can be observed that the maximum level of ASP and CCP between two waveforms depends upon the HFH sequence (H_k) i.e. frequency-firing sequence and the PN-code sequence (c_m). Thus, an optimization process is presented for the selection of DFCW-PNLC to minimize ASP level and CCP level of $\chi_v^{Multiuser}(\tau, f_d)$ and $\chi_W^{MIMO}(\tau, f_d)$.

The optimization process consists of two stages. In the first stage, following processes are performed:

- 1) Selection of two HFH sequences (H_k) for which the corresponding DFCW-LC, given by

$$\tilde{g}_k(t) = \sum_{i=1}^N g_k(t - (i-1)T) \quad (15)$$

provides minimum cross-correlation peak level is performed i.e. we need to minimize $\chi_{\tilde{g}_k, \tilde{g}_l}(\tau)$ where

$$\chi_{\tilde{g}_k, \tilde{g}_l}(\tau) = \int_{-\infty}^{\infty} \tilde{g}_k(t) \tilde{g}_l^*(t - \tau) dt \quad (16)$$

and $1 \leq k, l \leq p-1$.

- 2) Considering that a total of W waveforms are to be selected, select the remaining $W-2$ waveforms in an iterative process i.e choose the 3rd HFH sequence and corresponding waveform such that the total cross-correlation level remain as low as possible i.e. minimize

$$\chi_{\tilde{g}_l}(\tau) = \int_{-\infty}^{\infty} \sum_{w=1}^n \tilde{g}_{k_w}(t) \tilde{g}_l^*(t - \tau) dt \quad (17)$$

where $n = 2$, k_w denotes k th HFH sequence for w th waveform, given by $\tilde{g}_{k_w}(t)$, and $l \neq k_w$. Repeat the process (step 2) for the selection of the reaming waveforms.

Thus, first stage performs the optimization of the cross-correlation between DFCW-LC. In the second stage, following processes are performed:

- 1) Once the selection of discrete frequency-coding waveforms with linear chirp subpulses is completed, select any two optimized DFCW-LC (out of W) and for these two waveforms select two Gold code sequences (c_m) such that the two DFCW-PNLC, given by using (15) and (5) as

$$s_{k_w}^{m_w}(t) = \sum_{i=1}^N c_{m_w}(i) g_{k_w}(t - (i-1)T), \quad w = 1, 2 \quad (18)$$

provides minimum cross-correlation peak level i.e. using (18) and (6) minimize the cross-ambiguity function

$$\chi_{s_{k_1}^{m_1}, s_{k_2}^{m_2}}(\tau, f_d) = \int_{-\infty}^{\infty} s_{k_1}^{m_1}(t) s_{k_2}^{m_2*}(t - \tau) dt \quad (19)$$

considering $f_d = 0$ where $1 \leq m_w \leq p-1$.

- 2) After selection of two DFCW-PNLC, similar to the selection of HFH sequences in step 2 of first stage, encode the 3rd optimized DFCW-LC with PN-code such that overall cross-correlation peak level remains as low as possible i.e. minimize

$$\chi_{s_{k_3}^{m_3}}(\tau, f_d) = \int_{-\infty}^{\infty} \sum_{w=1}^n s_{k_w}^{m_w}(t) s_{k_3}^{m_3*}(t - \tau) dt \quad (20)$$

Repeat the process of step 2 for the selection of PN-code sequences of the reaming DFCW-LC.

IV. RESULTS AND PERFORMANCE ANALYSIS

Based on the optimization process, presented in Section III, a set of DFCW-PNLC is obtained. It is considered that the discrete frequency-coding waveforms can cover a total bandwidth (B_T) of 20 MHz. The design parameters considered for the construction of the proposed waveforms is given in Table I. For $W = 6$, the output of the optimization process provides the

TABLE I
DESIGN PARAMETERS FOR THE CONSTRUCTION OF THE PROPOSED DFCW-PNLC

Parameters	Values
Chirp subpulse bandwidth (B)	2 MHz
Duration of chirp subpulse (T)	400 ns
p	127
Frequency step (Δf)	$(B_T - B)/p$
Number of users or radar signals (W)	6
Doppler frequency span (f_d)	± 10 kHz

following selection of HFH sequences and code sequences as presented in Table II. It is worth to mention that the length of the Gold code sequences is $2^l - 1$ ($l = 2, 3, 4, 5, 6, 7$), and the number of subpulses used for the formation of the waveform is 126 (N). Thus, we have considered the Gold

TABLE II
HFH SEQUENCE NO. (k) AND PN-CODE SEQUENCE NO. (m) SELECTED THROUGH THE OPTIMIZATION PROCESS CONSIDERING $W = 6$

Radar signal No. (w)	HFH sequence No. - k (Corresponding HFH sequence - H_k)	PN-code sequence No. - m (Corresponding Gold code sequence - c_m)
1	9	13
2	23	55
3	47	26
4	67	69
5	73	102
6	46	96

code sequences of length 127 and the last element of the code sequence is dropped to match the length with the number of subpulses. Cross-correlation between two radar DFCW-PNLC $u_1(t)$ and $u_2(t)$, selected though the optimization process, is shown in Fig. 2(a) with maximin CCP level as -25.33 dB while for the autocorrelation of $u_1(t)$, maximum ASP level is calculated as -23.82 dB. Performance of autocorrelation and cross-correlation considering arbitrary k and m is shown

in Fig. 2(b) with maximin ASP and CCP level as -22.14 dB and -22.17 dB, respectively which clearly indicates that the proposed optimization process effectively minimizes the ASP and CCP levels.

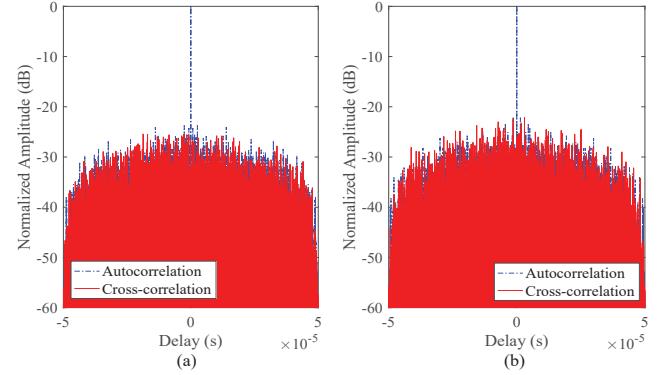


Fig. 2. (a) Autocorrelation and cross-correlation of two DFCW-PNLC $u_1(t)$ and $u_2(t)$, selected though the optimization process, where $u_1(t) = s_9^{13}(t)$ and $u_1(t) = s_2^{55}(t)$. (b) Autocorrelation and cross-correlation of two DFCW-PNLC chosen arbitrarily.

Performance of a radar user (considered user 3 i.e. $v = 3$) in a multiuser radar system, considering $W = 6$, is presented in Fig. 3 where the received signal is mixed with signals from all other 5 users. The output is shown for zero doppler scenario. The maximum sidelobe peak level of the correlated output is calculated as -16.44 dB when the signal of the desired user ($v = 3$) is present whereas if the receiver receives signals from other radar systems only, maximum CCP level is calculated as -17.13 dB. Performance of a multiuser radar system using

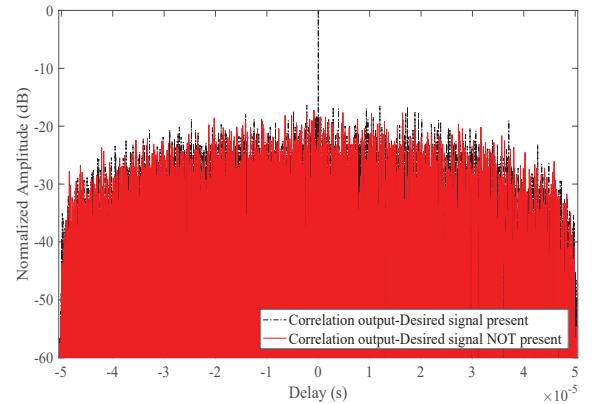


Fig. 3. Plot of multiuser radar ambiguity function for radar user 3 ($v = 3$) and considering $W = 6$.

waveforms designed according to different techniques [2], [3], [4], [5], [7], [8], considering the same sequence length (N), is compared with the proposed one and the corresponding maximin cross-correlation levels are listed in Table III. It can be observed that the proposed DFCW-PNLC provides a cross-correlation gain of 1.48 dB, 1.72 dB, 2.2 dB and 1.02 dB, 0.94 dB as compared to [8], [2], [3], [7] and [5], respectively.

TABLE III
COMPARISON OF MAXIMUM SIDELOBE PEAK LEVELS FOR MULTIUSER RADAR SYSTEM WITH $W = 6$

Waveform type	Max. sidelobe peak level (dB) (when signal of desired user is available)	Max. CCP level (dB) (when signal of desired user is not available)
Proposed DFCW-PNLC	-16.44	-17.13
DFCW-LC [8]	-15.01	-15.65
DFCW-FF [3], [4]	-14.37	-14.93
DFCW-FF with random phase [7]	-15.52	-16.11
Spread spectrum coded chirp [2]	-14.62	-15.41
DFPCW [5]	-15.57	-16.19

For a MIMO radar system, considering the number of transmitting and receiving antennas as 6, 6 optimized radar signals ($u_w(t)$) are transmitted through different antennas. The received composite signal can be analyzed through MIMO radar ambiguity function given by (13). The output correlated signal at zero doppler is shown in Fig. 4 with maximum sidelobe peak level is measured as -21.32 dB. In this context, pairwise

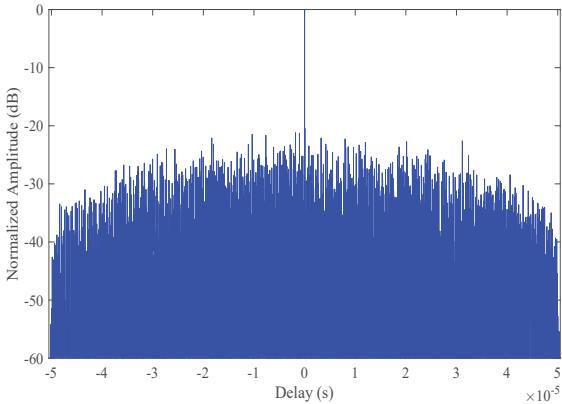


Fig. 4. Plot of MIMO radar ambiguity function considering $W = 6$.

TABLE IV
COMPARISON OF MAXIMUM SIDELOBE PEAK LEVELS FOR MIMO RADAR SYSTEM WITH $W = 6$

Waveform type	Max. sidelobe peak level (dB)
Proposed DFCW-PNLC	-21.32
DFCW-LC [8]	-19.64
DFCW-FF [3], [4]	-18.89
DFCW-FF with random phase [7]	-19.05
Spread spectrum coded chirp [2]	-17.73
DFPCW [5]	-18.92

cross-correlation between different transmitted DFCW-PNLC is calculated through simulation and maximum CCP level is found to be approximately -25.41 dB while maximum ASP

level is approximately -23.74 dB. The performance of the proposed waveforms for MIMO radar system is compared with the performance of the MIMO radar systems designed using different available radar waveforms such as DFCW-FF, DFCW-LC [2], [3], [4], [5], [7], [8] considering the same sequence length (N). The comparison output is listed in Table IV. It can be observed that combination of PN-code sequence with DFCW-LC and the selection of sequences through the proposed optimization process provides better performance in terms of maintaining lower sidelobe peak level as compared to other MIMO radar systems designed with different available waveforms.

Considering a Doppler variation of 10 kHz, Fig. 5 presents the gain loss due to the effect of Doppler frequency variation in the received signal. It can be found that MIMO radar system is more resilient as compared to multiuser radar system for the same Doppler variation, while the level of the sidelobe peaks remain approximately same during the variation.

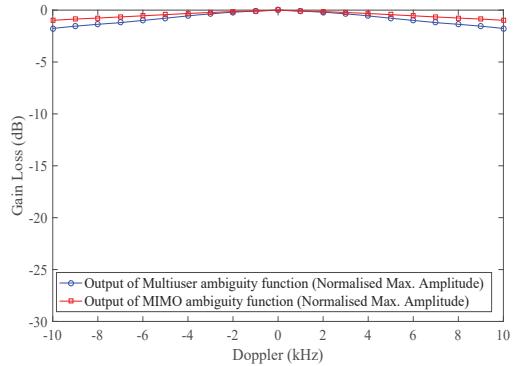


Fig. 5. Gain loss due to variation in Doppler frequency.

V. CONCLUSION

In this paper, design of discrete frequency-coding waveform with phase-coded linear chirp pulses is presented for multiuser and MIMO radar systems. HFH sequences are used to control the frequency-firing order of the waveforms. The waveform design process is discussed in details and the formulation of auto-ambiguity function and cross-ambiguity function, to analyze the performance of the waveforms, is presented. To improve the cross-correlation levels between a set of waveforms, an optimization process consisting of two stages is presented. Selection of the appropriate HFH sequences is performed in the first stage to optimize the cross-correlation between DFCW-LC, while the PN-code sequences are selected in such a way that the cross-correlation level between DFCW-PNLC become lower than the first stage and remains as low as possible in the second stage to obtain the final DFCW-PNLC. Simulation results show that the optimization process effectively minimizes the ASP and CCP levels of the waveforms. Performance of a multiuser and MIMO radar system considering 6 waveforms is presented. In terms of ASP and CCP level, performance of the multiuser and MIMO radar

systems designed using proposed waveforms provide better performance as compared to the multiuser and MIMO radar systems designed using existing DFCW which demonstrates the effectiveness of the proposed design technique and the optimization process.

REFERENCES

- [1] R. M. Davis, R. L. Fante, and R. P. Perry, "Phase-coded waveforms for radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 1, pp. 401–408, Jan. 2007.
- [2] S. J. Cheng, W. Q. Wang, and H. Z. Shao, "Spread spectrum-coded OFDM chirp waveform diversity design," *IEEE Sensors Journal*, vol. 15, no. 10, pp. 5694–5700, Oct. 2015.
- [3] H. Deng, "Discrete frequency-coding waveform design for netted radar systems," *IEEE Signal Process. Lett.*, vol. 11, no. 2, pp. 179–182, Feb. 2004.
- [4] W. Mehany, L. Jiao, and K. Hussien, "Orthogonal discrete frequency-coding waveform design based on modified genetic algorithm for MIMO-SAR," in *IEEE Conference on Industrial Electronics and Applications*, China, Jun. 2014, pp. 1082–1086.
- [5] G. Chang, X. Yu, and C. Yu, "Discrete frequency and phase coding waveform for MIMO radar," *Radio Engineering*, vol. 26, no. 3, pp. 835–841, Sep. 2017.
- [6] C. Y. Chen and P. P. Vaidyanathan, "MIMO radar ambiguity properties and optimization using frequency-hopping waveforms," *IEEE Trans. Signal Process.*, vol. 56, no. 12, pp. 5926–5936, Dec. 2008.
- [7] D. Deb, R. Bhattacharjee, and A. Vengadarajan, "Design of orthogonal waveforms for MIMO radar with phase randomised frequency hopping (FH) sequence," in *IEEE Conference on Electrical, Computer and Electronics Engineering*, India, Dec. 2016, pp. 235–238.
- [8] B. Liu, "Orthogonal discrete frequency-coding waveform set design with minimized autocorrelation sidelobes," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 45, no. 4, pp. 1650–1657, Oct. 2009.
- [9] H.A. Khan, Y. Zhang, C. Ji, C.J. Stevens, D.J. Edwards, and D. O'Brien, "Optimizing polyphase sequences for orthogonal netted radar," *IEEE Signal Process. Lett.*, vol. 13, no. 10, pp. 589–592, Oct. 2006.
- [10] C. E. Cook and M. Bernfeld, *Radar Signals - An Introduction to Theory and Application*, Academic Pr, 111 Fifth Avenue, New York, 1967.
- [11] X. Xia, J. Liu, and M. H. Lee, "A generalized hyperbolic frequency hopping sequences," in *International Symposium on Intelligent Signal Processing and Communication Systems*, South Korea, Nov. 2004, pp. 645–647.
- [12] S. Haykin and M. Moher, *Modern Wireless Communications*, Upper Saddle River: Pearson Education, 2004.

Saliency Guided Image Detail Enhancement

Sanjay Ghosh, Rituraj G. Gavaskar, and Kunal N. Chaudhury

Department of Electrical Engineering,

Indian Institute of Science,

Bangalore, India.

Email: {sanjayg, riturajg, kunal}@iisc.ac.in

Abstract—The use of visual saliency for perceptual enhancement of images has drawn significant attention. In this paper, we explore the idea of selectively enhancing salient regions of an image. Moreover, we develop an algorithm based on adaptive bilateral filtering for this purpose. In most of the filtering based methods, detail enhancement is performed by decomposing the image into base and detail layers; the detail layer is amplified and added back to the base layer to obtain the enhanced image. The decomposition is performed using edge-preserving smoothing such as bilateral filtering. The present novelty is that we use the saliency map to locally guide the smoothing (and the enhancement) action of the bilateral filter. The effectiveness of our proposal is demonstrated using visual results. In particular, our method does not suffer from gradient reversals and halo artifacts, and does not amplify fine details in non-salient regions that often appear as noise grains in the enhanced image. Moreover, if we choose to perform the filtering channelwise, then our method can be efficiently implemented using an existing fast algorithm.

Index Terms—Detail enhancement, saliency, bilateral filter, adaptation, fast algorithm.

I. INTRODUCTION

Detail enhancement is commonly used for improving the visual appearance of an image. This is typically done by increasing the local contrast and amplifying details. In most existing techniques, the latter is achieved by decomposing the image into base and detail layers [1]. The detail layer is then linearly amplified and added back to the base layer to get the enhanced image. The decomposition is performed using filtering [2], [3], [4], [5] or optimization [6], [7].

The human visual system cannot process all parts of an image at once [8]. It is biased towards focusing the gaze on *salient* regions, which stand out from the surroundings. While looking at a scene, we pay greater attention to these salient regions, while relatively ignoring the non-salient ones. Motivated by this observation, we propose to use visual saliency for detail enhancement. The objective is to boost the details in salient regions, while keeping the less salient regions unchanged (see Figure 1). This can be considered as an instance of *attention retargeting* [8]. The proposed technique is essentially built around two fundamental operations, namely, saliency detection and decomposition of the image into base and detail layers. The former assigns a numerical saliency value to each pixel in the image; pixels from salient regions are assigned larger

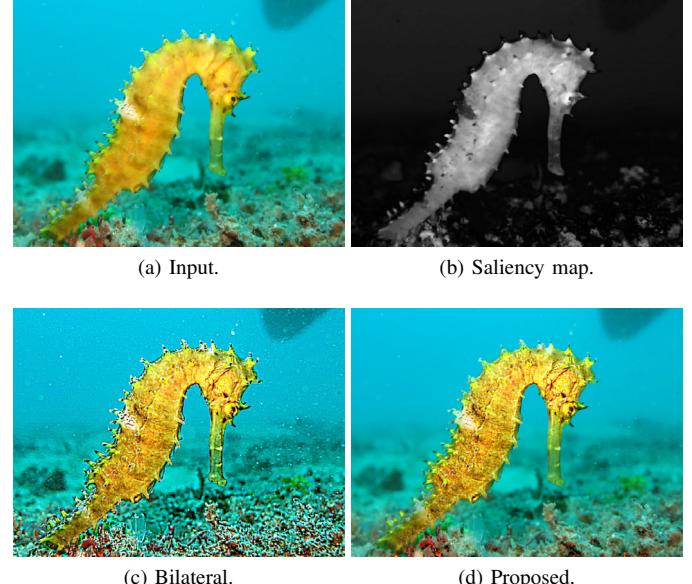


Fig. 1. Detail enhancement results using (c) standard and (d) saliency guided bilateral filtering. Notice the inadvertent amplification of fine details (appearing as noise grains) in non-salient regions in (c).

values. The bilateral filter [9] is commonly used for the base-detail decomposition. For our method, we propose to adapt the decomposition at each pixel based on its saliency value. In particular, this is achieved by guiding the smoothing action of the filter using the saliency map.

Research on saliency has gained momentum recently [10], [11], [12], [13], [14], [15]. Saliency information has been incorporated in applications such as visualization [10], image diffusion [11], retargeting [12], compression [13], tone mapping of HDR images [14], etc. The authors in [13] introduced a compression technique to minimize perceptual distortion in salient regions. Saliency was used in [12] to perform context-adaptive retargeting in the compressed domain. It was shown in [11] that the edge-preservation feature of the bilateral filter can be improved using saliency. To enhance the perception of volume data, a saliency-based enhancement operator was proposed in [10]. Saliency information was used in [14] to develop a local tone-mapping algorithm for displaying HDR images on devices with limited dynamic range. The saliency map is used to retain the fine details in a HDR image while reducing its dynamic range.

This work was partially supported by an EMR Grant SERB/F/6047/2016-2017 from the Department of Science and Technology, Government of India.

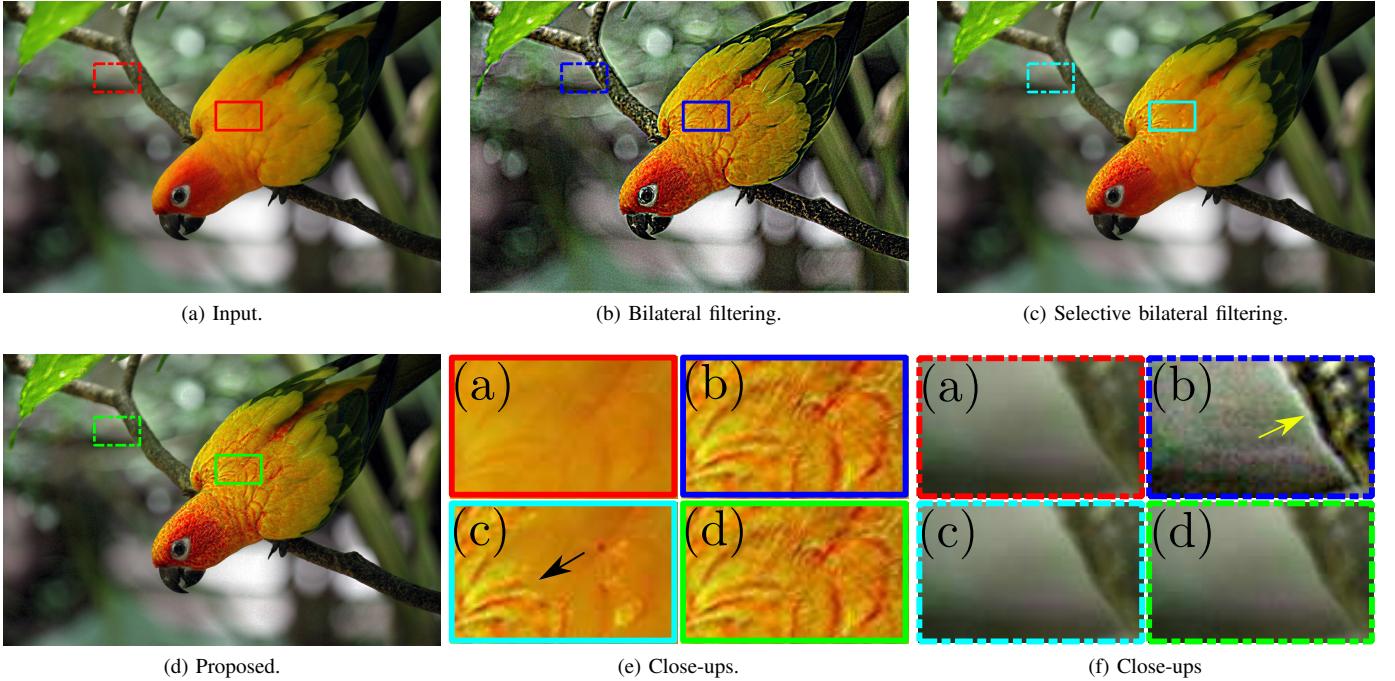


Fig. 2. Detail enhancement of (a) using three different methods. The base layer in (b) was obtained by bilateral filtering the entire image. In (c), only those pixels whose (normalized) saliency index is greater than $T = 0.5$ were processed to get the base layer. The result in (d) was obtained using adaptive bilateral filtering, where the amount of smoothing at each pixel is controlled via the saliency index. The enhancement factor is $K = 6$ in all cases. The other parameters are $\rho = 5$, $\sigma = 70$, and $\lambda = 10$ (see text for details). In Zoom 1, we can see discontinuities (one of which is marked with an arrow) arising from the *hard* assignment used in (c). On the other hand, in Zoom 2, we can see that standard bilateral filtering produces a halo along an edge (marked with an arrow) and inadvertently amplifies the noise grains.

The success of the above proposals motivated us to investigate the use of saliency for detail enhancement. In this regard, we note that content-aware detail enhancement was proposed in [7], where the idea is to selectively amplify the image gradient based on the image content. An optimization framework based on L_0 -minimization was recently used for this purpose in [6]. To the best of our knowledge, the present idea of adaptively regulating the bilateral filter using the saliency map is new. An added advantage of this adaptive version of the bilateral filter is that we can use an existing fast algorithm [16] to speed up the filtering (performed on a channel-by-channel basis).

The rest of the paper is organized as follows. The technical details of the proposed method are presented in Section II. In particular, we discuss a standard framework for detail enhancement, an existing method for saliency detection, and the proposed method for guiding the bilateral filtering based on saliency. In Section III, we report visual results and compare with a recent method. We conclude the paper with a discussion in Section IV.

II. SALIENCY GUIDED ENHANCEMENT

A. Enhancement using bilateral filtering

Suppose we wish to enhance the details of a RGB color image $\mathbf{f}(x)$. As mentioned earlier, a standard technique [1] is to decompose the image into base and detail layers $\mathbf{h}(x)$ and $\mathbf{d}(x)$ such that

$$\mathbf{f}(x) = \mathbf{h}(x) + \mathbf{d}(x). \quad (1)$$

Ideally, the base layer contains structures such as edges, coarse textures, etc. The detail layer captures fine textures and the contrast. The enhanced image $\mathbf{g}(x)$ is obtained by amplifying the detail layer by some factor $K > 1$, which is then added back to the base layer:

$$\mathbf{g}(x) = \mathbf{h}(x) + K\mathbf{d}(x). \quad (2)$$

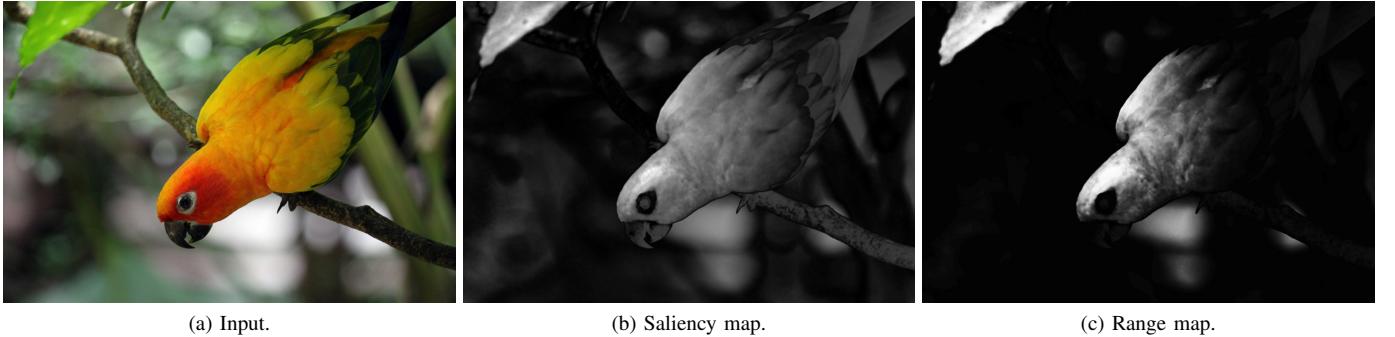
A higher value of K leads to greater enhancement; however, a large K can also introduce artifacts in the output image.

In filtering-based methods [2], [3], [4], [5], the base layer is obtained using edge-preserving smoothing such as bilateral filtering [9]. We recall that the bilateral filter uses a range kernel along with a spatial kernel for edge-preserving smoothing. In particular, we use a Gaussian for the former and a box for the latter. Both these choices are quite common [17]. We use a box spatial kernel for computational simplicity. The base layer is computed using

$$\mathbf{h}(x) = \frac{1}{\eta(x)} \sum_{y \in \Omega} \exp \left(-\frac{\|\mathbf{f}(x-y) - \mathbf{f}(x)\|^2}{2\sigma^2} \right) \mathbf{f}(x-y), \quad (3)$$

where $\Omega = [-\rho, \rho]^2$ is a window centered at $(0, 0)$, $\|\cdot\|$ denotes the Euclidean norm, and

$$\eta(x) = \sum_{y \in \Omega} \exp \left(-\frac{\|\mathbf{f}(x-y) - \mathbf{f}(x)\|^2}{2\sigma^2} \right). \quad (4)$$



(a) Input.

(b) Saliency map.

(c) Range map.

Fig. 3. Saliency map for the image in (a) computed using (5). Also shown is the corresponding range map given by (6). The former is normalized to $[0, 1]$, while the latter is on a scale of $[0, 70]$.

We note that this is the color version of the bilateral filter. Alternatively the grayscale version of the bilateral filter may be used to separately filter each channel of a color image.

Following (1), the detail layer is set as

$$\mathbf{d}(x) = \mathbf{f}(x) - \mathbf{h}(x).$$

The parameter σ in (3) and (4) controls the amount of smoothing induced by the filter. When σ is large, the bilateral filter behaves like a box filter; at the other extreme, its smoothing action is turned off when σ is small [9].

The above mechanism enhances the entire image, including the less salient regions. However, it is often desirable to keep the less salient regions unchanged. For example, by enhancing details in non-salient regions, we might obtain noise grains in the enhanced output. To illustrate this point, we refer the reader to Figure 2(b), which is obtained by enhancing Figure 2(a) using a bilateral filter ($\sigma = 70$ on a scale of $[0, 255]$). We can see fine grains appearing in Figure 2(b). In general, the enhanced image in Figure 2(b) appears noisy compared to the input image. A possible solution is to restrict the edge-preserving smoothing (and hence the enhancement) to essentially the salient regions. In fact, the problem in Figure 2(b) is satisfactorily resolved using the present proposal (see Figure 2(d)), where we use saliency to softly turn off the enhancement process in regions with relatively less saliency.

B. Saliency computation

Before turning to our main proposal, we discuss the method used to build the saliency map. A saliency algorithm assigns a scalar value to each pixel, where pixels having higher values are considered to be more salient. The saliency map is constructed by maximizing the mutual information of feature distributions in [18], and using center-surround feature distances in [19]. The approach in [20] is based on finding regions which imply unique frequencies in the Fourier domain. A contrast-based approach is used in [21], [22]. The method in [23] is based on the idea that salient regions are distinctive with respect to both their local and global surroundings. Co-occurrence histogram is used in [24], which uses both global pixel occurrences and local co-occurrence of pixel pairs within a neighborhood. A model built on the distinctness of patch patterns is proposed in

[25], and an optimization method based on both sparsity and distinctness is presented in [26].

Among the above approaches, we found that [20] is well suited to our purpose due to two reasons. Firstly, the saliency maps have sharp edges between salient and non-salient regions. This prevents halo artifacts from appearing in the enhanced image, as evident from the results shown in Section III. Secondly, the cost of computing the saliency map is relatively low. In particular, the saliency map $S(x)$ is computed as

$$S(x) = \|\mathbf{f}_G(x) - \boldsymbol{\mu}\|^2, \quad (5)$$

where $\boldsymbol{\mu}$ is the mean pixel value of the input image in CIE-Lab space. $\mathbf{f}_G(x)$ is the Gaussian filtering of $\mathbf{f}(x)$; the filtering is performed channelwise in the RGB space and the filtered result is transformed to the CIE-Lab space. This is used to remove fine textures, noise and compression artifacts, if any, from the input image. We will use (5) for saliency computation for our purpose. The saliency map obtained using (5) can assume any non-negative value. For reasons that will be evident shortly, we linearly normalize the saliency map to $[0, 1]$, which we continue to denote using $S(x)$. The saliency map computed using (5) is shown in Figure 3.

C. Adaptive bilateral filtering

A first approach towards saliency driven enhancement would be to enhance just the salient pixels. To do so, we first need to partition the image into salient and non-salient regions. This can be done by fixing some threshold T , and identifying a pixel as being salient if $S(x) \geq T$. Although simple and intuitive, *hard* assignment can produce discontinuities in the output, e.g., when there is a sudden transition from a salient to a non-salient region (see Figure 2(c)).

A more reliable approach is to smoothly control the smoothing parameter σ in (3) in tune with the saliency values. In particular, we propose to use the following *range map*

$$\sigma(x) = \Psi_{\lambda, T}(S(x)), \quad (6)$$

where

$$\Psi_{\lambda, T}(t) = \frac{\sigma}{1 + \exp(-\lambda(t - T))} \quad (7)$$

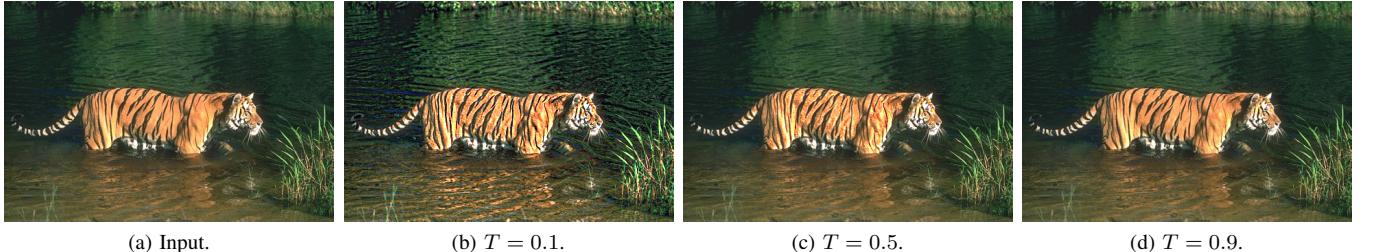


Fig. 4. Enhancement results for different T in (7). The other parameters are $K = 4$, $\rho = 5$, $\sigma = 80$, and $\lambda = 20$. Almost the entire image is enhanced if T is small, while there is no enhancement when T is close to 1. Setting T to be somewhere in between, we can selectively enhance the salient regions.

is a scaled and shifted sigmoid. The parameters σ , T and λ correspond to the scale, shift, and slope of the sigmoid respectively. Notice that (7) approaches a step function when λ is large. This essentially corresponds to the hard thresholding scheme discussed previously. We found that satisfactory results are obtained when $\lambda \in [10, 30]$. For such values, (6) avoids discontinuities by gradually transitioning from high to low saliency regions. The range map computed using (6) for $\sigma = 70$, $T = 0.5$, and $\lambda = 10$ is shown in Figure 3.

In summary, our proposal is to generate the base layer using

$$\mathbf{h}(x) = \frac{1}{\eta(x)} \sum_{y \in \Omega} \exp \left(-\frac{\|\mathbf{f}(x-y) - \mathbf{f}(x)\|^2}{2\sigma(x)^2} \right) \mathbf{f}(x-y), \quad (8)$$

where

$$\eta(x) = \sum_{y \in \Omega} \exp \left(-\frac{\|\mathbf{f}(x-y) - \mathbf{f}(x)\|^2}{2\sigma(x)^2} \right).$$

It follows from (6) that $\sigma(x)$ is small (resp. large) when $S(x)$ is small (resp. large). This ensures that the smoothing action in (8) is turned off softly in regions with low saliency. As a result, $\mathbf{h}(x) \approx \mathbf{f}(x)$ and $\mathbf{d}(x) \approx 0$; hence, $\mathbf{g}(x) \approx \mathbf{f}(x)$ from (2). On the other hand, in regions with large saliency, $\sigma(x) \approx \sigma$, where the proposed method performs detail enhancement as given by (2) and (3).

We note that there are papers where the range kernel is adapted pixelwise for denoising and to reduce compression and registration artifacts [27], [28], [29]. However, the present idea of adapting it with saliency for detail enhancement is new.

III. RESULTS AND ANALYSIS

The performance of the proposed method is determined by the choice of σ and T . The parameter T acts as a soft threshold for discriminating high-saliency regions from low-saliency ones. Setting $T \approx 0$ results in the enhancement of almost the entire image. In contrast, setting T close to 1 gives a near identical image to the input, with very little detail enhancement. This is illustrated in Figure 4. The images in Figures 4(b), (c) and (d) are enhanced versions of the input image in Figure 4(a) for $T \in \{0.1, 0.5, 0.9\}$. We typically set $T \in [0.2, 0.5]$ for our experiments. Moreover, we observed that setting $\sigma \in [50, 80]$ generally gives good enhancement results. Needless to say, the optimal parameter settings depend on the input image. In a typical application, the user would be expected to change T and σ to explore different enhancement results.

In Figure 2, we compare our result with those obtained by bilateral filtering the entire image and just those pixels where $S(x) \geq 0.5$. Notice that our method does not introduce discontinuities or unpleasant grainy artifacts and halos in regions with relatively less saliency (see description in the caption). In summary, the proposed method combines the desirable properties of global and selective bilateral filtering, while suppressing their unwanted influences.

We note that the use of saliency for detail enhancement has not been explored in the literature. The closest match is the content-aware enhancement method from [7]. To demonstrate the effectiveness of our proposal, we compare it with standard bilateral filtering and [7]. The results are shown in Figures 5 and 6. The detail enhancement in salient regions is clearly evident in both figures. Notice that the results obtained using [7] suffer from color distortions. In contrast, the images obtained using our method does not exhibit color distortions or artifacts arising from gradient reversal [3].

We also compare our method with [7] using the quality metrics BIQI [30], NIQE [31], and SQMS [32]. For the no-reference quality metrics BIQI and NIQE, a smaller value represents higher quality. On the other hand, SQMS is a full-reference image quality metric which is particularly designed for saliency-guided quality measurement; a higher value of SQMS indicates better quality. These metrics are mentioned in the captions of Figures 5 and 6 in the order (BIQI, NIQE, SQMS). Notice that our method achieves superior performance in terms of all three metrics. The reason for the latter is that there is no significant enhancement near an edge (which is indeed a desirable property [7]) in our method. In particular, the saliency map takes low values near edges (see Fig. 3). In fact, the saliency map exhibits a sharp transition in saliency values between two sides of an edge. This phenomenon is due to the efficacy of the saliency detection algorithm in [20]. As a result, $\sigma(x)$ is small, which results in virtually no smoothing at those pixels and the output is almost identical to the input.

Finally, we note that if we choose to process the color channels separately, then we can use the fast algorithm in [16] to accelerate the filtering. We provide an example in Figure 7, where we have performed detail enhancement using the color bilateral filter in (8) as well as the fast algorithm in [16] to generate the base layer. Notice that while the two results are visually similar, the latter is more than 10 times faster.

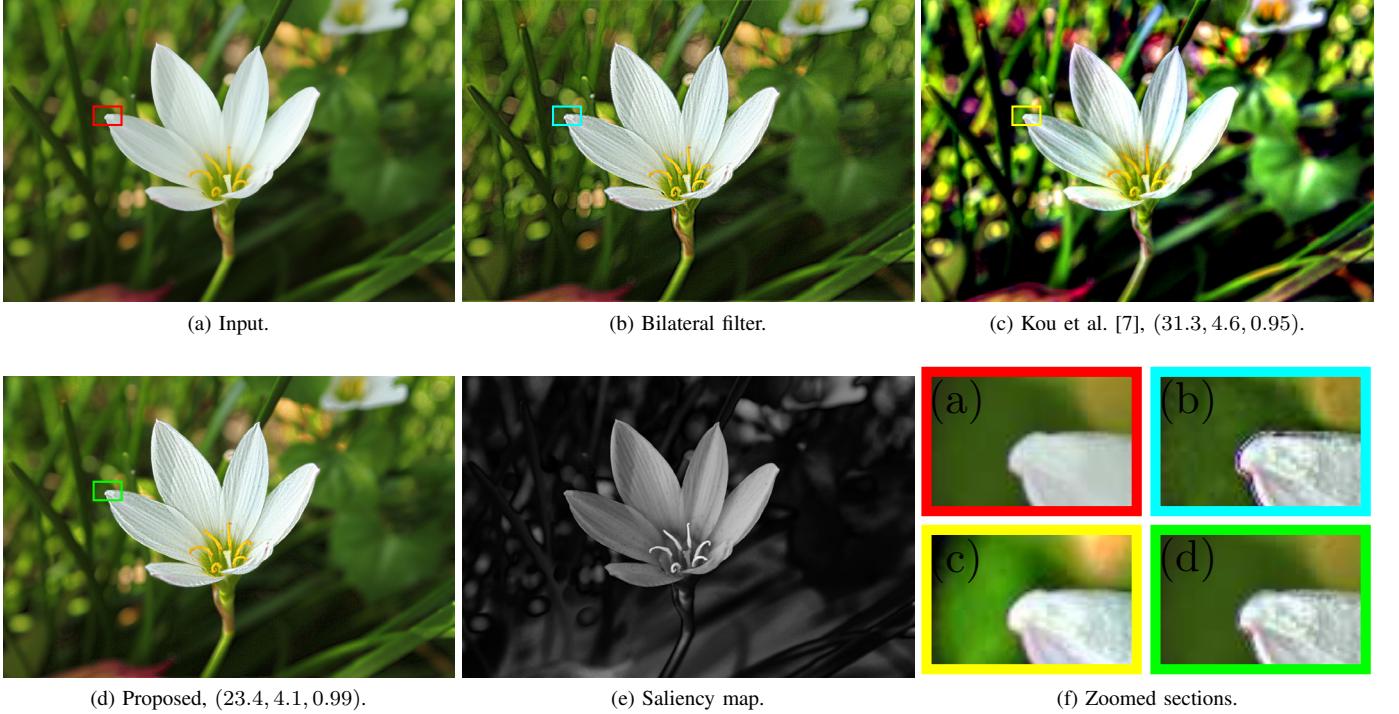


Fig. 5. Comparison of detail enhancement results. The parameters are $\rho = 5$, $\sigma = 60$, $\lambda = 20$, $T = 0.5$, and $K = 5$. We can clearly see color distortions in (c) and gradient reversals in (d).

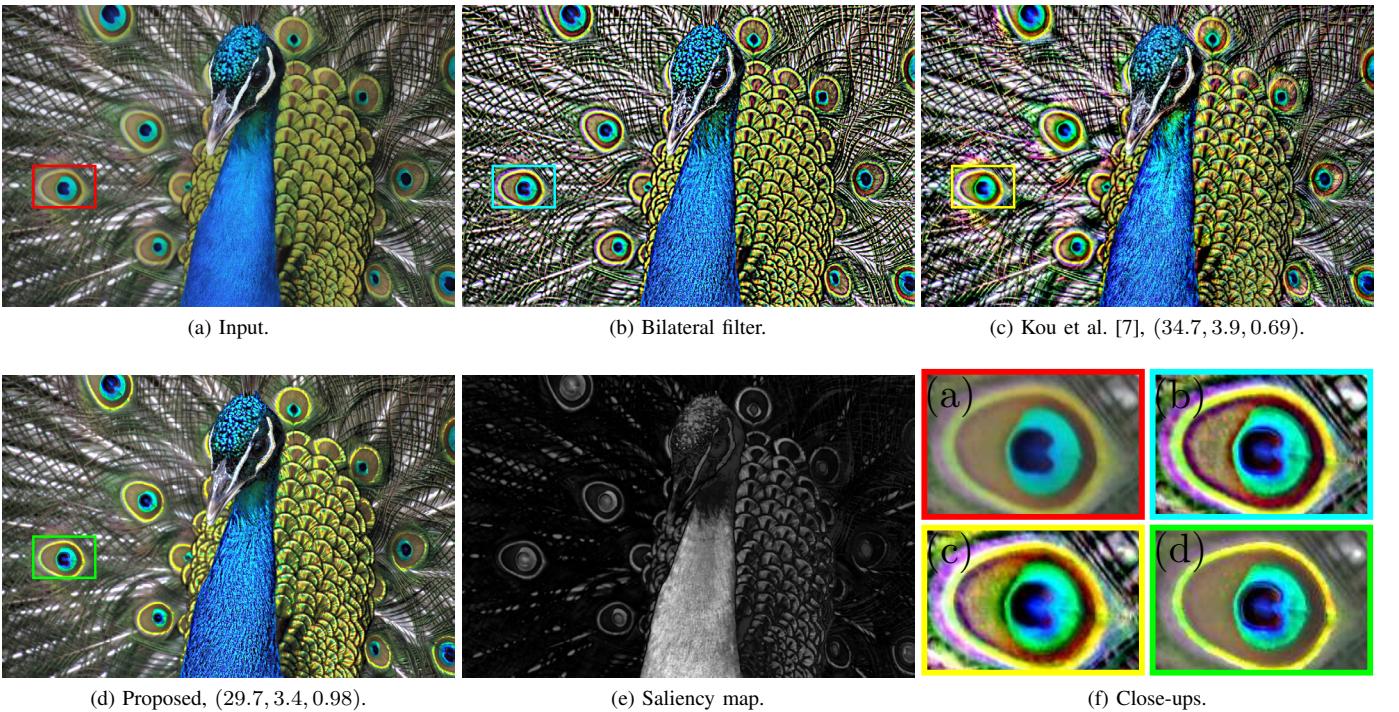


Fig. 6. Comparison of detail enhancement results. The parameters are $\rho = 5$, $\sigma = 70$, $\lambda = 30$, $T = 0.2$, and $K = 6$. We see significant color distortions in (b) and (c) compared to our result (see zoomed sections). Notice that new colors have appeared throughout (c).



Fig. 7. Results obtained by performing the bilateral filtering in the the combined color space (c) and on a channel-by-channel basis (d). The channelwise filtering is performed using the fast algorithm in [16]. The timings for (c) and (d) are 13.6 and 1.2 seconds. The (BIQI, NIQE, SQMS) values are also reported.

IV. CONCLUSION

We explored the idea of using a saliency-guided bilateral filter for detail enhancement. The key novelty was the use of saliency index to softly control the inhibitory action of the range kernel and, in effect, the smoothing induced by the filter. The method comes with tunable parameters for subjectively enhancing images based on user preference. Importantly, we demonstrated that our method does not introduce undesirable visual artifacts such as color distortion and halos in the enhanced image. We also showed that if each channel is filtered independently, then the proposed method admits an efficient implementation.

REFERENCES

- [1] Z. Farbman, R. Fattal, D. Lischinski, and R. Szeliski, “Edge-preserving decompositions for multi-scale tone and detail manipulation,” *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 67:1–67:10, 2008.
- [2] R. Fattal, M. Agrawala, and S. Rusinkiewicz, “Multiscale shape and detail enhancement from multi-light image collections,” *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 51:1–51:9, 2007.
- [3] K. He, J. Sun, and X. Tang, “Guided image filtering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.
- [4] M. G. Mozerov and J. van de Weijer, “Global color sparseness and a local statistics prior for fast bilateral filtering,” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5842–5853, 2015.
- [5] Q. Zeng, H. Qin, H. Leng, Xiang X. Yan, J. Li, and Huixin H. Zhou, “Adaptive detail enhancement for infrared image based on bilateral filter,” *AOPC 2015: Image Processing and Analysis*, vol. 9675, pp. 96752N, 2015.
- [6] L. Xu, C. Lu, Y. Xu, and J. Jia, “Image smoothing via l0 gradient minimization,” *ACM Transactions on Graphics*, vol. 30, no. 6, pp. 174:1–174:12, 2011.
- [7] F. Kou, W. Chen, Z. Li, and C. Wen, “Content adaptive image detail enhancement,” *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 211–215, 2015.
- [8] J. Tsotsos, “Analyzing vision at the complexity level,” *Behavioral and Brain Sciences*, vol. 13, no. 3, pp. 423–445, 1990.
- [9] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” *Proc. IEEE International Conference on Computer Vision*, pp. 839–846, 1998.
- [10] Y. Kim and A. Varshney, “Saliency-guided enhancement for volume visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, 2006.
- [11] J. Xie, P. Heng, and M. Shah, “Image diffusion using saliency bilateral filter,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 6, pp. 768–771, 2008.
- [12] Y. Fang, Z. Chen, W. Lin, and C. W. Lin, “Saliency detection in the compressed domain for adaptive image retargeting,” *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, 2012.
- [13] S. Li, M. Xu, Y. Ren, and Z. Wang, “Closed-form optimization on saliency-guided image compression for hevc-msp,” *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 155–170, 2018.
- [14] Z. Li and J. Zheng, “Visual-saliency-based tone mapping for high dynamic range images,” *IEEE Transactions on Industrial Electronics*, vol. 61, no. 12, pp. 7076–7082, 2014.
- [15] R. Mechrez, E. Shechtman, and L. Zelnik-Manor, “Saliency driven image manipulation,” *IEEE Winter Conference on Applications of Computer Vision*, pp. 1368–1376, 2018.
- [16] R. G. Gavaskar and K. N. Chaudhury, “Fast adaptive bilateral filtering,” *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 779–790, Feb 2019.
- [17] F. Porikli, “Constant time $O(1)$ bilateral filtering,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
- [18] D. Gao and N. Vasconcelos, “Bottom-up saliency is a discriminant process,” *Proc. IEEE International Conference on Computer Vision*, pp. 1–6, 2007.
- [19] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, “Salient region detection and segmentation,” *Proc. International Conference on Computer Vision Systems*, pp. 66–75, 2008.
- [20] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, “Frequency-tuned salient region detection,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, 2009.
- [21] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, “Global contrast based salient region detection,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–416, 2011.
- [22] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, 2012.
- [23] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.
- [24] S. Lu, C. Tan, and J-H. Lim, “Robust and efficient saliency modeling from image co-occurrence histograms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 195–201, 2014.
- [25] R. Margolin, A. Tal, and L. Zelnik-Manor, “What makes a patch distinct?,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1139–1146, 2013.
- [26] Y. Luo, P. Wang, W. Zhu, and H. Qiao, “Sparse-distinctive saliency detection,” *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1378–1382, 2015.
- [27] B. Zhang and J. P. Allebach, “Adaptive bilateral filter for sharpness enhancement and noise removal,” *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 664–678, 2008.
- [28] M. Zhang and B. K. Gunturk, “Compression artifact reduction with adaptive bilateral filtering,” *Proc. SPIE Visual Communications and Image Processing*, vol. 7257, 2009.
- [29] S. Mangiat and J. Gibson, “Spatially adaptive filtering for registration artifact removal in hdr video,” *Proc. IEEE International Conference on Image Processing*, pp. 1317–1320, 2011.
- [30] Anush Krishna Moorthy and Alan Conrad Bovik, “A two-step framework for constructing blind image quality indices,” *IEEE Signal Processing Letters*, vol. 17, no. 5, pp. 513–516, 2010.
- [31] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, “Making a ‘completely blind’ image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [32] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang, “Saliency-guided quality assessment of screen content images,” *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

Caching Partial Files for Content Delivery

V S Ch Lakshmi Narayana
IIT Bombay
164076006@iitb.ac.in

Sambhav Jain
IIT Bombay
sambhavjain975@gmail.com

Sharayu Moharir
IIT Bombay
sharayum@ee.iitb.ac.in

Abstract—Numerous empirical studies have shown that users of video-on-demand platforms do not always watch videos in their entirety. A direct consequence of this is that not all parts of a video are equally popular. Motivated by this, we explore the benefits of dividing files into smaller segments for caching. We treat incoming requests as requests for segments of files and propose a Markovian request model which captures the time-correlation in requests. We characterize the fundamental limit on the performance of caching policies which only cache full files. Next, we propose and analyze the performance of policies which cache partial files. Using this, we characterize the potential for improvement in performance due to caching partial files and analyze its dependence on various system parameters like cache size and the popularity profile of the files being cached.

I. INTRODUCTION

Recent years have witnessed a steep increase in the popularity of video-on-demand (VoD) services. It is estimated that 73% of the current Internet traffic can be attributed to video traffic and this fraction is expected to increase to 82% over the next five years [1]. One way of meeting this ever increasing demand without overloading the available network resources is to cache content close to the end-users. Delivering content to users via geographically co-located caches has the dual positive effect of reducing the load on the network backbone, thus reducing the overall bandwidth consumption of the network and improving the quality of service to the end-users.

Caching techniques for VoD services have received a lot of attention in the last few years [2]–[7]. One aspect that has not received sufficient attention is the fact that videos are often not viewed in their entirety. Empirical studies have shown that on average, only 60% of a video is watched by viewers [8], [9]. A direct consequence of this is that not all parts of a video are equally popular and this motivates dividing videos into chunks/segments for caching [8], [10]–[13].

Most of the work on caching techniques for VoD services caches full files and models the request arrival process as an i.i.d. process [2]–[7]. This assumption is not valid if videos are divided into segments for storage as a viewer watching a part of a video is quite likely to request the next segment of the same file next. This time-correlation in requests requires new request models and caching techniques and is the focus of this work. We use a Markov Chain to capture this time-correlation.

We consider a system consisting of a central server which stores the entire catalog of contents on offer, and a local cache deployed close to the end-user with limited storage resources.

This work was supported in part by the Bharti Centre for Communication at IIT Bombay and a seed grant from IIT Bombay .

978-1-5386-9286-8/19/\$31.00 © 2019 IEEE

The algorithmic challenge is to determine which file segments to cache locally in order to minimize the fraction of requests that have to be routed to the central server for service.

The key contributions of this work can be summarized as follows.

- We model the request arrival process as a Markov Chain to capture the time-correlation in incoming requests.
- We first focus on static storage policies. For any static storage policy, the contents of the cache are not modified on request arrivals. We characterize the optimal static storage policy in the class of policies that are constrained to store full files and the optimal storage policy when files are allowed to be split into segments for caching. We evaluate the benefits of dividing files into segments for caching via analysis and simulations.
- Next, we focus on dynamic caching policies where if the requested content is not available in the local cache, it is fetched as stored in the cache to serve the request. We characterize the fundamental limit on the performance of any caching policy for our setting. We propose a caching policy which stores segments of files and compare its performance with the fundamental limit. In addition, we compare the performance of our policy with two other popular caching policies via simulations.

A. Related Work

Several empirical studies have found that most users do not watch videos in their entirety [8], [9], [11], [14], [15]. In [8], the authors state that, on average, only 60% of a video is watched by viewers. In [9], the authors observe that 60% of the Youtube videos are watched for less than 20% of their duration. Moreover, 80% of Youtube videos are interrupted due to lack of user interests [15] .

Potential advantages of storing segments of files instead of caching complete files have been explored in [8], [10]–[13], [16], [17]. In [10] a segmentation scheme is proposed when video requests follow the Independent Reference Model (IRM). The proposed segmentation scheme is shown to be optimal using via analysis and empirical evaluation. In [11], the problem of caching partial files is formulated as a mixed integer problem. Based on user viewing patterns, more precisely video retention rate of users, [8] proves that chunking of videos is helpful and proposes an algorithm called chunk-LRU for caching partial files. In this work, user request process is modeled using the IRM and users abandon videos after a random portion of time. [17] proposes three file segmentation schemes, namely, Fixed segmentation, Pyramid segmentation and Skyscraper segmentation. In the Fixed segmentation scheme, each file is divided into segments of equal size.

In Pyramid segmentation, the size of the segments grows exponentially as we go deeper into a video. The size of the segments grows slowly under Skyscraper segmentation scheme as compared to that in Pyramid segmentation. These three schemes lead to better byte hit ratio when compared to full-video caching approach and among these three schemes, the Pyramid segmentation scheme performs the best. In [12], the authors propose a decentralized caching scheme called Inter-chunk Popularity based Edge first Caching (IPEC) which exploits the benefits of storing partial files. In [13], the partial caching problem is studied for the performance metric of bandwidth consumption.

The key modeling difference between these works and our work is that we model the request arrival process as a Markov Chain. This was also the case in [18], [19], however, in [18], [19], the arrival process is modeled as a Markovian processes to capture the time-correlation in requests coming from a user as a result of recommendations made by the VoD platform.

II. SETTING

A. Servers and Storage

We study a system consisting of a central server and a local cache. The central server has sufficient storage capacity to store all the files being offered and can serve all the requests that are routed to it. The local cache has limited storage capacity, i.e., the cache can store up to S of the N files in the catalog, where typically, $S < N$.

B. Request Process

The content catalog consists of N files of equal size, each divided into J segments of equal size. The request arrival process is Markovian with the following structure. After viewing a specific segment of a file, the user requests for the next segment of the same file with some probability and requests for the first segment of a fresh file otherwise. After watching the last segment of a file, the next request is always for the first segment of a fresh file. Formally, the request process is a random walk on a directed weighted graph consisting of $(N \times J + 1)$ states defined as follows.

Assumption 1 (Markovian Request Process):

- Let $v_{i,j}$ represent the j^{th} segment of the i^{th} file and v_0 be a dummy node which represents the state of not watching a video.
- We construct a weighted graph $G(V, E)$, where

$$V = \{v_{i,j}, \text{ for } 1 \leq i \leq N, 1 \leq j \leq J\} \cup \{v_0\}.$$

$$E = \{(v_{i,j}, v_{i,j+1}), \text{ for } 1 \leq i \leq N, 1 \leq j \leq J-1\}$$

$$\cup \{(v_{i,j}, v_0), \text{ for } 1 \leq i \leq N, 1 \leq j \leq J-1\}$$

$$\cup \{(v_{i,J}, v_0), \text{ for } 1 \leq i \leq N\}$$

$$\cup \{(v_0, v_{i,1}), \text{ for } 1 \leq i \leq N\}, \text{ and}$$

$$w_e = \begin{cases} p_i^{j+1} & \text{for } e \in \{(v_{i,j}, v_{i,j+1}), 1 \leq i \leq N, \\ & 1 \leq j \leq J-1\}, \\ p_i^{j0} = 1 - p_i^{j+1} & \text{for } e \in \{(v_{i,j}, v_0), 1 \leq i \leq N, \\ & 1 \leq j \leq J-1\}, \\ 1 & \text{for } e \in \{(v_{i,J}, v_0), \text{ for } 1 \leq i \leq N\}, \\ p_i & \text{for } e \in \{(v_0, v_{i,1}), \text{ for } 1 \leq i \leq N\}. \end{cases}$$

where w_e is the weight of the directed edge e .

- The request process is a random walk on the directed weighted graph $G(V, E)$, with the probability of making a transition from a node to one of its neighbors being proportional to the corresponding edge weights.

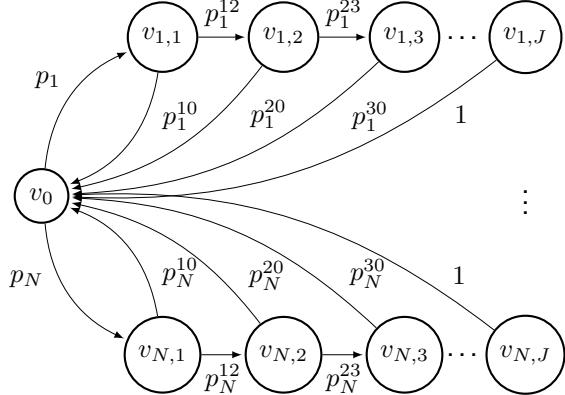


Fig. 1: Markov Chain for the catalog size of N files and each divided into J segments.

We use the following definition in the rest of this paper.

Definition 1: Request Cycle – Sequence of segments requested between two consecutive visits to the dummy node.

C. Service Model

Each incoming request is served by the local cache if the requested file is available in the cache. We refer to this event as a ‘hit’. If the requested file is not present in the cache, the request is served by fetching the requested file from the central server. This event is referred to as a ‘miss’.

D. Goal

The goal is to design caching policies which maximize the expected number of hits per request cycle. This metric can be used to derive the standard metric of hit-rate using the renewal theorem [20].

III. STATIC STORAGE POLICIES

We first focus on a class of policies called Static Storage policies. For any policy belonging to this class, the cache contents are fixed before the request arrival process begins and cannot be modified thereafter.

Definition 2: Popularity of a Segment ($q_{i,j}$) – Under Assumption 1, the popularity of the segment $v_{i,j}$ is defined as the probability of it being requested in a request cycle, i.e.,

$$q_{i,j} = \begin{cases} p_i & \text{for } j = 1, \\ p_i \times \prod_{k=1}^{j-1} p_i^{kk+1} & \text{otherwise.} \end{cases}$$

Next, we define and study a policy which caches full files.

Definition 3: Store Most Popular Files (SMPF) – The SMPF policy stores the S most popular files, i.e., S files with the S highest $\sum_{j=1}^J q_{i,j}$ values (Definition 2), breaking ties arbitrarily.

Lemma 1: Under Assumption 1, in the class of static storage policies that are constrained to store full files, i.e., either all

segments of a file are cached or none at all, the SMPF policy is optimal.

Proof: Let $x_{i,j}$ be an indicator random variable which indicates if the j^{th} segment of the i^{th} file belongs to a request cycle. Under Assumption 1, the probability of a segment $v_{i,j}$ being requested in a request cycle is given by $q_{i,j}$. Therefore, $\mathbb{E}[x_{i,j}] = q_{i,j}$.

For a static storage policy P constrained to store full files, let the set of files cached be denoted by \mathcal{S}_P . For this policy, let the expected number of hits in a request cycle be denoted by h_P . Under Assumption 1, we have that

$$h_P = \sum_{i \in \mathcal{S}_P} \sum_{j=1}^J \mathbb{E}[x_{i,j}] = \sum_{i \in \mathcal{S}_P} \sum_{j=1}^J q_{i,j}.$$

It follows that storing the S files with the highest $\sum_{j=1}^J q_{i,j}$ values maximizes the expected number of hits, thus making the SMPF policy optimal in the class of static storage policies constrained to store full files. ■

Next, we define and study a caching policy which divides files into segments for caching.

Definition 4: Store Most Popular Segments (SMPS): The SMPS policy stores the SJ most popular segments i.e., the SJ segments with maximum $q_{i,j}$ values (Definition 2), where i and j are file index and segment index respectively, breaking ties arbitrarily.

Lemma 2: Under Assumption 1, in the class of static storage policies that are allowed to store parts of files (segments), the SMPS policy is optimal.

Proof: Let $x_{i,j}$ be an indicator random variable which indicates if the j^{th} segment of the i^{th} file belongs to a request cycle.

For a static storage policy P which is allowed to store partial files, i.e., segments, let the set of segments cached be denoted by \mathcal{S}_P . For this policy, let the expected number of hits in a request cycle be denoted by h_P . Under Assumption 1, we have that

$$h_P = \sum_{v_{i,j} \in \mathcal{S}_P} \mathbb{E}[x_{i,j}] = \sum_{v_{i,j} \in \mathcal{S}_P} q_{i,j}.$$

It follows that h_P is maximized by caching segments in the set \mathcal{S}^* , where,

$$\mathcal{S}^* = \arg \max_{S: |\mathcal{S}| \leq SJ} \sum_{v_{i,j} \in \mathcal{S}} q_{i,j}.$$

It follows that storing the SJ most popular segments maximizes the expected number of hits in a request cycle, thus making the SMPS policy optimal in the class of static storage policies allowed to store file segments. ■

We now evaluate the performance of these policies for following simple popularity profile.

Assumption 2 (Two-Class Popularity Model):

The arrival process satisfies Assumptions 1 with the following parameter values.

Files are divided into two classes, namely, Class A and Class B such that there are $N_A = N^\alpha$, $\alpha < 1$ files in Class A (say

files 1 to N_A) and the remaining $N_B = N - N_A$ files in Class B. The popularity of all files in a class is equal, i.e.,

$$p_i = \begin{cases} p_A & \text{if File } i \text{ is in Class A,} \\ p_B & \text{otherwise.} \end{cases}$$

Motivated by the observation that in most popular video on demand services, a small set of files account for a large fraction of requests, we focus on the case where $p_A \gg p_B$. More specifically, for a given positive constant $r < 1$, $p_A = r/N^\alpha$ and $p_B = (1 - r)/(N - N^\alpha)$. In addition,

$$w_e = \begin{cases} p & \text{for } e \in \{(v_{i,j}, v_{i,j+1}), 1 \leq i \leq N, \\ & \quad 1 \leq j \leq J-1\}, \\ 1-p & \text{for } e \in \{(v_{i,j}, v_0), 1 \leq i \leq N, \\ & \quad 1 \leq j \leq J-1\}, \\ 1 & \text{for } e \in \{(v_{i,J}, v_0), \text{ for } 1 \leq i \leq N\}, \\ p_i & \text{for } e \in \{(v_0, v_{i,1}), \text{ for } 1 \leq i \leq N\}. \end{cases}$$

Our next result compares the performance of the SMPS policy with the performance of an optimal static storage policy that is constrained to store full files for the Two-Class Popularity Model, i.e., SMPF.

Proposition 1: Consider a system with a cache that can store N^γ files for $\gamma < 1$. Let the arrival process satisfy Assumption 2 and h_{SMPF} and h_{SMPS} be the expected number of hits per request cycle for the SMPF and SMPS policies respectively. Then, we have that,

- (a) If $\alpha > \gamma$ (small cache), $\frac{h_{\text{SMPS}}}{h_{\text{SMPF}}} = \frac{J(1-p)}{(1-p^J)} > 1$.
- (b) If $\alpha < \gamma$ (large cache), $\lim_{N \rightarrow \infty} \frac{h_{\text{SMPS}}}{h_{\text{SMPF}}} = 1$.

Proof: Let $w = \sum_{i=1}^J p^{i-1}$.

- (a) If $\alpha > \gamma$, $h_{\text{SMPF}} = N^\gamma p_A w$, and $h_{\text{SMPS}} = N^\gamma p_A J$.
- (b) If $\alpha < \gamma$, for N large enough,

$$\begin{aligned} h_{\text{SMPF}} &= N^\alpha p_A w + (N^\gamma - N^\alpha)p_B w \\ h_{\text{SMPS}} &= N^\alpha p_A w + (N^\gamma - N^\alpha)p_B J, \\ \implies \frac{h_{\text{SMPS}}}{h_{\text{SMPF}}} &= \frac{J(1-p)}{(1-p^J)} > 1. \end{aligned}$$

We thus conclude that for the Two-Class Popularity Model (Assumption 2), the gain from storing partial files is higher for small caches. ■

A. Simulation Results

We now present simulation results for the case where content popularity follows the Zipf distribution. This is motivated by the fact that numerous empirical studies have shown that the popularity profile for many VoD services follows the Zipf distribution [21]. Typical values of β lie between 0.6 and 2.

Assumption 3 (Zipf Popularity Model):

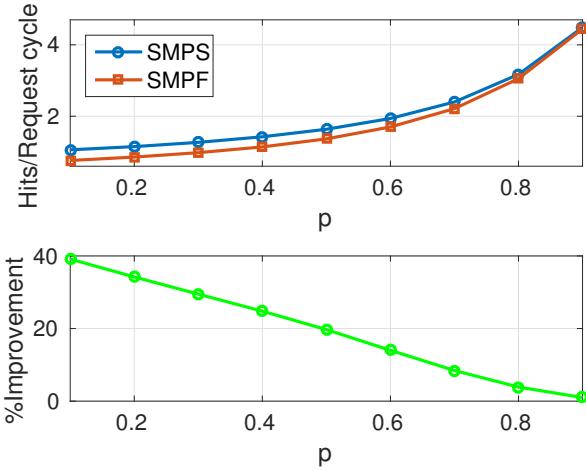


Fig. 2: Comparison of number of hits per request cycle for the SMPF and SMPS policies for the Zipf Popularity Model (Assumption 3) for different p values.

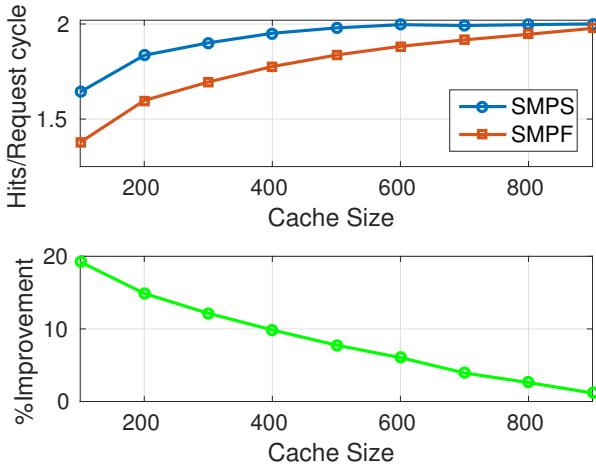


Fig. 3: Comparison of number of hits per request cycle for the SMPF and SMPS policies for the Zipf Popularity Model (Assumption 3) for different cache sizes.

The arrival process satisfies Assumption 1 with $p_i \propto i^{-\beta}$, where $\beta > 0$ is the Zipf parameter. The remaining edge weights are as in Assumption 2.

In the simulation results presented in this section, we consider a catalog of 100 files, each divided into 10 segments. Figure 2 compares the performance of the SMPF and SMPS policies for different values of p for a cache size of 100 segments and Zipf parameter $\beta = 1.2$. The SMPS policy outperforms the SMPF policy for all values considered. We see that the difference between the performances of these two policies is a decreasing function of p . This is expected since as the value of p increases, the probability that the entire file is requested increases, thus making SMPF more effective.

Figure 3 compares the performance of the SMPF and SMPS policies for different values of cache size for $p = 0.5$ and $\beta = 1.2$. We see that number of hits per request cycle increases with increase in cache size since more files/segments can be

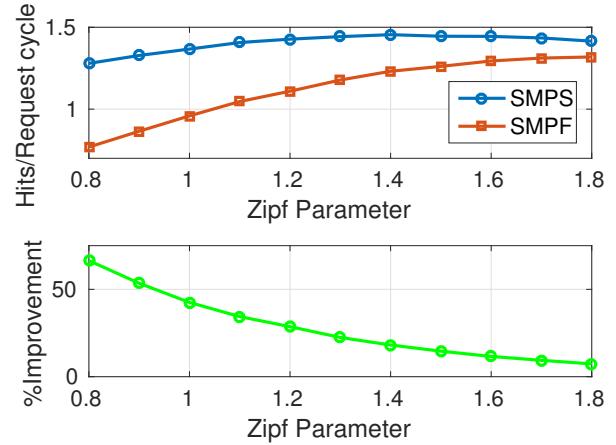


Fig. 4: Comparison of number of hits per request cycle for the SMPS and SMPF policies using Zipf Popularity Model (Assumption 3) for different values of the Zipf Parameter.

cached. The SMPS policy outperforms the SMPF policy for all values considered and performance gap between these two policies decreases with increase in cache size.

Figure 4 compares the performance of the SMPF and SMPS policies for different values of the Zipf parameter β . Note that a higher value of β leads to a more lopsided popularity distribution. Here, the cache can store 100 segments and value of p is 0.5. The performance of both the policies improves as β increases. This is because as β increases, the cached files/segments account for a larger fraction of requests. The SMPS policy outperforms the SMPF policy for all values of β considered, however, the difference in their performance decreases as β increases.

IV. CACHING POLICIES

In traditional caching systems, on a miss, the requested file is fetched from the central server and stored in the cache to serve the request. In this section we look at three such caching policies and compare their performance for the request arrival process defined in Section II.

We now characterize an upper bound on the expected number of hits in a request chain for any caching policy which satisfies the following constraints. The contents of the cache can only be modified on a miss. In addition, on a miss, only the requested content can be fetched and stored on the cache by removing any one of the currently cached contents.

Theorem 1: Let \mathcal{S}_{SJ} be the set of the SJ most popular segments and h^* be the expected number of hits in a request chain for the optimal caching policy. For a system consisting of a cache that can store S full files or SJ segments and if the request arrival process satisfies Assumption 1, we have that,

$$h^* \leq \sum_{v_{i,j} \in \mathcal{S}_{SJ}} q_{i,j}.$$

Proof: Among the segments requested in a request chain, requests for only those segments that are in the cache at the beginning of the request chain can result in hits. This is direct consequence of the fact that under Assumption 1, each segment is requested at most once in a request chain and therefore,

segments fetched due to misses in a chain do not lead to hits in the same chain. Given this, the proof of $h^* \leq \sum_{v_{i,j} \in \mathcal{S}_{SJ}} q_{i,j}$ follows on the same lines as that of Lemma 2. ■

A. Remove Least Popular Segment (RLPS) policy

The first caching policy we consider is the Remove Least Popular Segment (RLPS) policy. As the name suggest, on a miss, when a segment is fetched from the central server and the cache is full, this policy replaces the least popular segment currently in the cache with the fetched segment and serves the request. Refer to Algorithm 1 for a formal definition.

Algorithm 1: Remove Least Popular Segment (RLPS)

- 1 Initialize: Set of stored segments \mathcal{S}_{RLPS} = the set of $S \times J$ segments with the highest $q_{i,j}$ values
 - 2 On receiving a request for segment s :
 - 3 if $s \notin \mathcal{S}_{RLPS}$ then
 - 4 fetch segment s from the central server
 - 5 find $s_{\text{least-popular}} = \arg \min_{v_{i,j} \in \mathcal{S}_{RLPS}} q_{i,j}$, breaking ties arbitrarily
 - 6 $\mathcal{S}_{RLPS} = (\mathcal{S}_{RLPS} \setminus \{s_{\text{least-popular}}\}) \cup \{s\}$
 - 7 end
 - 8 Serve the request, goto 2
-

Our next result characterizes the performance of the RLPS policy.

Theorem 2: Let \mathcal{S}_n be the set of the n most popular segments and h_{RLPS} be the expected number of hits in a request chain under the RLPS policy. For a system consisting of a cache that can store S full files or SJ segments and if the request arrival process satisfies Assumption 1, we have that,

$$h_{RLPS} \geq \sum_{v_{i,j} \in \mathcal{S}_{SJ-1}} q_{i,j}.$$

Proof: By the definition of the RLPS policy, the set of $SJ - 1$ most popular segments are always available in the cache. Therefore, $h_{RLPS} \geq \sum_{v_{i,j} \in \mathcal{S}_{SJ-1}} \mathbb{E}[x_{i,j}]$, where $x_{i,j}$ is an indicator random variable which indicates if the j^{th} segment of the i^{th} file belongs to a request chain. Under Assumption 1, $\mathbb{E}[x_{i,j}] = q_{i,j}$, thus proving the result. ■

B. Least Recently Used (LRU) policy

We now study the popular Least Recently Used caching policy. This is a widely used and studied caching policy. In this section, we evaluate the performance of LRU when the request arrival process satisfies Assumption 3. Refer to Algorithm 2 for a formal definition.

We now provide a lower bound on the number of hits per request cycle for the LRU policy for the Zipf Popularity Model (Assumption 3). We omit the proof due to lack of space.

Theorem 3: Let h_{LRU} be the expected number of hits in a request cycle in LRU policy. For a system consisting of a cache that can store SJ segments and if the request process satisfies Assumption 3, we have that

$$h_{LRU} \geq \sum_{\forall v_{i,j} \in P} q_{i,j} (1 - o(1)),$$

Algorithm 2: Least Recently Used (LRU)

- 1 Let the set of stored segments be \mathcal{S}_{LRU}
 - 2 Initialize: $\mathcal{S}_{LRU} = \emptyset$
 - 3 On receiving a request for segment s :
 - 4 if $s \notin \mathcal{S}_{LRU}$ then
 - 5 fetch segment s from the central server
 - 6 find $s_{\text{least-recently-used}} =$ the least recently used segment in the cache
 - 7 $\mathcal{S}_{LRU} = (\mathcal{S}_{LRU} \setminus \{s_{\text{least-recently-used}}\}) \cup \{s\}$
 - 8 end
 - 9 Serve the request, goto 3
-

$$\text{where } P = \left\{ v_{i,j} : q_{i,j} \geq \frac{(\log SJ)^{4\beta}}{(SJ)^\beta} \right\}.$$

C. Markov Paging Policy

The third caching policy we study is Markov Paging [22], known to be 2-optimal for any Markovian arrival process. The key idea behind the Markov Paging algorithm is to use the commute time between pair of nodes in the Markov chain to make caching decisions. Refer to [22] for the details.

D. Simulation Results

In this section we compare the performance of Markov paging, LRU and RLPS policies for the Zipf Popularity Model (Assumption 3). Along with these three policies, we also plot the upper bounds on number of hits per request cycle when full files are stored (UB Full Files) and when segmentation is allowed (UB Segments).

For the Zipf Popularity model, we consider a catalogue of $N = 100$ files, each file is divided into $J = 10$ segments and the cache can store 100 segments. We compare the performance of Markov paging, LRU, and RLPS policies for different values of β for $p = 0.5$ can be seen in Figure 5. With increase in β , the popularity profile of the files becomes lopsided and the cached segments account for a larger fraction of incoming requests. Therefore as β increases, the number of hits per request cycle increases. Here RLPS performs better among the three policies and its performance is close to the upper bound when segmentation is allowed. LRU and Markov paging policies performs better than all policies which store full files.

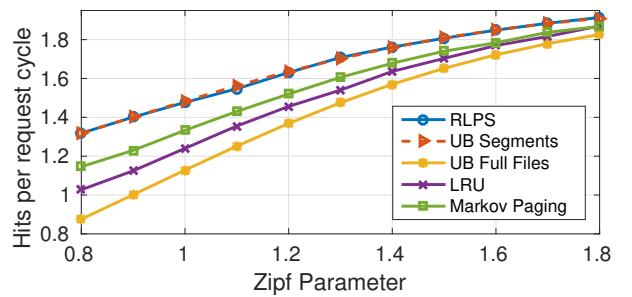


Fig. 5: Comparison of number of hits per request cycle for the RLPS, LRU and Markov Paging policies for the Zipf Popularity Model (Assumption 3) for different values of the Zipf Parameter.

Figure 6 compares the performance of Markov paging, LRU and RLPS policies for different values of p for Zipf parameter $\beta = 1.2$. With an increase in p , the number of hits per request cycle increases for all the three policies. From Figure 6 we see that RLPS policy outperforms the other two dynamic policies and its performance is close to the upper bound when segmentation is allowed.

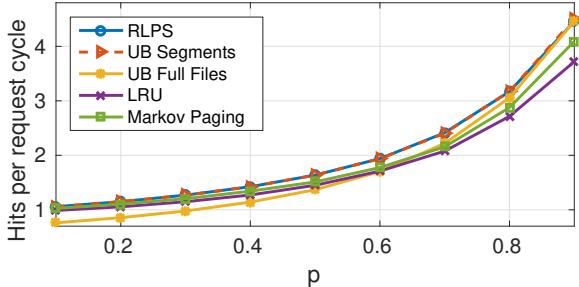


Fig. 6: Comparison of number of hits per request cycle for the RLPS, LRU and Markov Paging policies for the Zipf Popularity Model (Assumption 3) for different values of p .

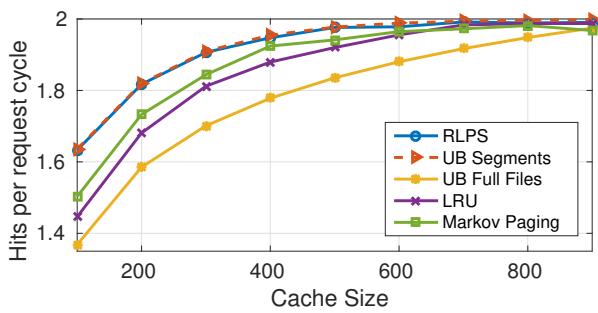


Fig. 7: Comparison of number of hits per request cycle for the RLPS, LRU and Markov Paging policies for the Zipf Popularity Model (Assumption 3) for different cache sizes.

Figure 7 compares the performance of the Markov paging, LRU and RLPS policies for different values of cache size for $p = 0.5$ and $\beta = 1.2$. We see that number of hits per request cycle increases with increase in cache size since more files/segments can be cached. From Figure 7 we see that RLPS policy outperforms the other two dynamic policies and its performance is close to upper bound when segmentation is allowed. The difference between the performance of RLPS and the upper bound for caching full files decreases as the cache size increases.

V. CONCLUSIONS AND FUTURE WORK

We focus on a single cache system in the setting where the popularity of segments of a file is non-uniform. We propose a Markovian request arrival process for this setting. We study both static storage and caching policies. For both cases, we first prove a fundamental upper limit on the performance of any policy that is constrained to store full files. We then propose and analyze the performance of policies which cache parts of files. We conclude that simple storage/caching policies which store partial files significantly outperform policies which

store entire files. In addition, we observe that the potential for improvement is higher when the cache is small and when the popularity is comparable across files.

Generalizing the results to the setting where the request arrivals form a renewal process and the setting where file popularity is known only within certain estimation error are potential directions of further study.

REFERENCES

- [1] Cisco VNI, "Cisco visual networking index: Forecast and methodology 2016–2021.(2017)," 2017.
- [2] S. Moharir and N. Karamchandani, "Content replication in large distributed caches," in *Communication Systems and Networks (COMSNETS), 2017 9th International Conference on*. IEEE, 2017, pp. 128–135.
- [3] V. Shah and G. De Veciana, "High-performance centralized content delivery infrastructure: models and asymptotics," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1674–1687, 2015.
- [4] V. Shah and G. de Veciana, "Performance evaluation and asymptotics for content delivery networks," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2014, pp. 2607–2615.
- [5] M. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.
- [6] S. Moharir, J. Ghaderi, S. Sanghavi, and S. Shakkottai, "Serving content with unknown demand: the high-dimensional regime," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1. ACM, 2014, pp. 435–447.
- [7] N. Karamchandani, U. Niesen, M. Maddah-Ali, and S. Diggavi, "Hierarchical coded caching," in *Information Theory (ISIT), 2014 IEEE International Symposium on*. IEEE, 2014, pp. 2142–2146.
- [8] G. P. Lorenzo Maggi, Lazaros Gkatzikis and J. Leguay, "Adapting caching to audience retention rate: Which video chunk to store?" 2015.
- [9] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: a view from the edge," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 15–28.
- [10] L. Wang, S. Bayhan, and J. Kangasharju, "Optimal chunking and partial caching in information-centric networks," 2015.
- [11] Hwang, D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, V. Misra, K. K. Ramakrishnan, and D. F. Swayne, "Leveraging video viewing patterns for optimal content placement," 2012.
- [12] S. H. Lim, Y. B. Ko, G. H. Jung, J. Kim, and M. W. Jang, "Inter-chunk popularity-based edge-first caching in content-centric networking," 2014.
- [13] J. Yu and C. T. Chou, "A dynamic caching algorithm based on internal popularity distribution of streaming media," 2014.
- [14] K.-W. Hwang, V. Gopalakrishnan, R. Jana, S. Lee, V. Misra, and K. Ramakrishnan, "Abandonment and its impact on p2p vod streaming," in *Peer-to-Peer Computing (P2P), 2013 IEEE Thirteenth International Conference on*. IEEE, 2013, pp. 1–10.
- [15] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "Youtube everywhere: Impact of device and infrastructure synergies on user experience," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 345–360.
- [16] S. Seny, J. Rexfordz, and D. Towsley, "Proxy prefix caching for multimedia streams," 1999.
- [17] K. L. Wu, H. S. Yu, and J. L. Wolf, "Segmentation of multimedia streams for proxy caching," 2004.
- [18] S. Gupta and S. Moharir, "Effect of recommendations on serving content with unknown demand," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2017, p. 35.
- [19] ———, "Request patterns and caching for vod services with recommendation systems," in *Communication Systems and Networks (COMSNETS), 2017 9th International Conference on*. IEEE, 2017, pp. 31–38.
- [20] W. L. Smith, "Renewal theory and its ramifications," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 243–302, 1958.
- [21] Y. Liu, F. Li, L. Guo, B. Shen, S. Chen, and Y. Lan, "Measurement and analysis of an internet streaming service to mobile devices," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 11, pp. 2240–2250, 2013.
- [22] A. R. Karlin, S. J. Phillips, and P. Raghavan, "Markov paging," 1997.

The HTTP/2 Server Push and Its Implications on Mobile Web Quality of Experience

Hema Kumar Yarnagula and Venkatesh Tamarapalli

Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, Guwahati, India

Email: {h.yarnagula, t.venkat}@iitg.ac.in

Abstract—In recent years, an unprecedented growth in the usage of mobile devices for web browsing poses a challenge for the service providers to assure the user-perceived quality. In the context of web quality of experience (QoE), quality perception is mostly dominated by the page load time (PLT). HTTP/2 protocol, with the server push feature, promises to address the design limitations of HTTP/1.1 that inhibit optimal web performance. However, it remains largely unclear if HTTP/2 can really improve web QoE for mobile browsing.

In this paper, we experimentally investigate the web QoE with HTTP/2. We assess the web QoE for several popular websites on a controlled testbed emulated with real 4G/LTE and 3G network traces. Our experiments investigate the impact of both network latency and packet loss ratio on the mobile web QoE. The results clearly show 24% improvement in the PLT, on an average, with HTTP/2 over mobile networks. However, we identify that HTTP/2 with server push is necessarily not the fail-safe solution for improving mobile web QoE under all conditions. We noticed that HTTP/2 loads the web pages slower than HTTP/1.1 when the network packet loss ratio is more than 2%. Our study could be used as the basis to derive a set of guidelines on the usage of the HTTP/2 server push to improve the end-user web QoE, especially in mobile devices.

I. INTRODUCTION

In the last two decades, the rise in adoption of mobile devices such as, smartphones and tablets has resulted in an explosive mobile data traffic growth. According to the statistics reported in Cisco VNI Forecast Update, there has been a remarkable 18-fold growth in the mobile data traffic over the past five years and around 81 percent of this is from the smartphones (including phablets) [1]. As this trend continues, it not only poses a formidable challenge for the network service provider(s) to systematically increase the network capacity but also to ensure an acceptable quality of experience (QoE) to the end-user(s). According to the ITU-T Rec.P.10 [2], the QoE is defined as the overall user's perception of the acceptability of an application or service. In the recent times, the research literature is mostly dominated by QoE assessment for multimedia services. There have also been some studies on the quality assessment of web browsing, popularly termed *web QoE* where, the quality perception is mostly dominated by the page load time (PLT). Larger the PLT, more dissatisfied is the end-user. Several studies on web QoE have been carried out in the literature that focused mainly on detecting inefficiencies with HTTP persistent connections [3], analyzing the ill-effects of the website complexity on user experience [4], the role of

mobile networks on web QoE [5], and study of the metrics and tools for web QoE assessment [6].

Over the years, it is noticed that the websites are growing larger, both in terms of the number of resource requests per page and the average web page size. According to [7], the average transfer size for Alexa top 1000 websites in 2016, is around 2 MB and on an average there are about hundred unique resource requests for each website. Though the Hypertext Transfer Protocol (HTTP/1.1) is the de-facto protocol used for the web page delivery, it was not designed to handle such large number of object requests. The Internet Engineering Task Force (IETF) in 2012, decided to address the well-known inefficiencies of HTTP/1.1 with a new version of HTTP, namely HTTP/2 [8], the first draft of which was adopted from Google's SPDY protocol [9]. Subsequently it was finalized as RFC 7540 in May 2015 with the major objective being reducing the PLT.

The HTTP/2 protocol not only retains the basic syntax and semantics of its predecessor but also provides several new features such as server push, request and response multiplexing, and header compression. The server push feature in HTTP/2 allows the server to push multiple objects to the client without explicit client request(s). Despite its potential, it remains largely unclear if the server push feature can really yield perceivable web QoE improvement for mobile web browsing. Motivated by this aspect, we experimentally investigate the impact of both HTTP/1.1 and HTTP/2 equipped with server push on mobile web QoE. More precisely, the contributions/findings of this paper are summarized as follows:

- 1) We conduct controlled web QoE assessment of several popular websites under the dynamics of mobile network characteristics (e.g, bandwidth variation, loss, and latency) to explore the implications of HTTP/2 server push on PLT. We use both 4G/LTE and 3G network traces in a testbed to emulate the mobile network conditions for our experiments. Our results demonstrate that mobile browsing with HTTP/2 server push leads to about 24% reduction in PLT under both 4G and 3G network conditions compared to HTTP/1.1.
- 2) In addition, we also investigated the impact of network latency and packet loss ratio on mobile web QoE. Our results demonstrate that increasing network latency significantly increases the PLT with both HTTP/1.1 and HTTP/2. However, push feature in HTTP/2 saves half an RTT per request to outperform HTTP/1.1 in terms of the PLT. Our findings give the indication that HTTP/2

server push can not be the straightforward solution to optimize the web QoE, unless used properly.

- 3) We study the complex relationship between different network parameters and TCP connection overhead for HTTP/1.1. Our results reveal that websites with JavaScript files of large size suffer from higher TCP connection overhead. Furthermore, our results demonstrate a notable increase in the number of TCP connections established for HTTP/1.1 with increasing network latency and PLR.

The rest of this paper is structured as follows. We present a brief introduction to HTTP/1.1 and HTTP/2, and discuss the related work on web QoE in Section II. Subsequently, we describe our experimental setup in detail and the procedure for the website selection in Section III. We present the experimental results and our observations on mobile web QoE in Section IV and conclude the paper in Section V.

II. BACKGROUND AND RELATED WORK

Over the time, the web pages have grown richer and also introduce complexities both in terms of the number of resource requests per page and total transfer size. Figure 1 illustrates the evolution of the average number of object requests per page and the average page transfer size, for mobile devices during the past four years [7]. From this figure, it can be seen that on an average, eighty requests are used to download a single web page of approximately 1600 KB in size. Similar trends are expected to follow in the years to come. However, the HTTP/1.1 protocol, that carries most of the web traffic these days, was not designed to effectively deal with a large number of requests/responses per page, limiting its performance.

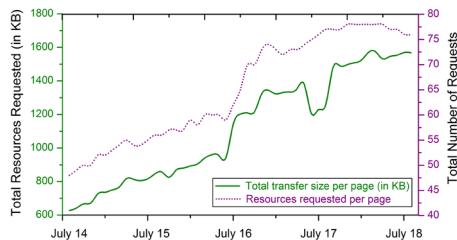


Fig. 1. Evolution of the average page size and number of object requests per page generated from mobile devices between July 2014 - July 2018

A. Design Limitations of HTTP/1.1

Head-of-line blocking (HOL blocking): Often the HTTP/1.1, with a persistent TCP connection, fetch several objects sequentially and it rarely makes a single object request. The total time required to load the page roughly turns to be the number of objects multiplied by the RTT between the client and server. The pipelining feature does allow the client to dispatch a group of requests over a single TCP connection at once. But, this feature is still implemented using a first-in first-out (FIFO) queue at the server. This ensures that the client receive the responses sequentially in the requested order. Any delay in processing a request from the queue (i.e., either a large file at the head-of-line which takes time or packet

loss) will eventually block the subsequent responses. This performance-limiting phenomenon is often called *head-of-line blocking* (HOL blocking). In addition, pipelining feature is disabled by default in all the major browsers which makes injudicious use of the limited resources in devices such as smartphones.

Fat request and response headers: The HTTP message headers vary in size roughly from 200 bytes to over 2 KB. With applications using more cookies and other features could increase the header to a few kilobytes in size. The HTTP/1.1 protocol does not provide any mechanism to compress the message headers. The network latency could pile up quickly just to send the resource requests, especially in mobile networks, where the up-link bandwidth is often constrained by the network provider.

B. HTTP/2

Several new features were introduced in HTTP/2 while retaining the basic syntax and semantics of HTTP/1.1. Among all these features, the server push is an important one. When a client makes a request for a web page, it is quite likely that the client will also make a handful of other requests for the dependencies such as images, style sheets and JavaScript files. In HTTP/1.1, the server waits for the client request(s) and only then serves the response(s). However, in HTTP/2, upon receiving a client request, the server push feature allows the server to predict all the associated page dependencies (e.g., CSS, JS and image files) with an assumption that these will be required in the near future. These are proactively pushed along with the response to the original client request. To accomplish this, the HTTP/2 server initiates the push stream by sending a PUSH_PROMISE frame on an existing client-initiated stream. The client can either accept or decline the PUSH_PROMISE stream. If the client has unexpired resource in its cache, pushing the same again is undesirable, particularly in cellular networks where the customers are charged on the volume of transferred data.

C. Related Work

The authors of [10] conducted large-scale experiments to test HTTP/2 performance and its adoption. They reported in the year 2016, that half of the websites using HTTP/2 implemented HTTP/1 optimization techniques like domain sharding (splitting web page resources across multiple servers), which results in establishing multiple TCP connections per page. In [11], a limited study with a virtual Ethernet interface, showed that merely switching to a newer protocol version could prove detrimental and is not enough to achieve significant reduction in PLT. It was suggested that configuring the HTTP/2 server intelligently and designing the websites keeping the HTTP/2 features in mind could give noticeable performance improvement. Furthermore, several web QoE studies for HTTP/1.1 (see e.g., [12], [13], [14]) primarily focus on mapping PLT to user-perceived quality estimates and investigated time-dependent properties of PLT. The work in [15] and [16] studied the energy consumption of HTTP/2

as compared to its predecessor for both mobile users and web servers, respectively. The authors also observed that HTTP/2 reduces the energy consumption for both mobile devices and web servers while improving the performance in networks with higher RTTs. Although, the metrics such as PLT provide both quantitative and qualitative insight on web QoE, a recent subjective study of users on desktop, clearly showed that this metric is not strongly correlated with the actual user opinion score [17]. While this could be considered as a limitation of our study, conducting a full-fledged subjective evaluation on mobile devices is practically hard. We leave the subjective evaluation of QoE for mobile browsing as our future work.

We conclude from the literature survey that it is difficult to derive direct conclusions on the impact of HTTP/2 on web QoE for mobile devices, as there are a wide range of rather conflicting results. To the best of our knowledge no work has been carried out to investigate the implications of HTTP/2 server push on mobile web QoE. This motivates us to take up the QoE assessment of HTTP/2 for web browsing on mobile devices under a controlled trace-driven testbed. Furthermore, we also investigate the individual impact of variations in both network latency and PLR on mobile web QoE.

III. EXPERIMENTAL METHODOLOGY

In this section, we provide the details about different components of the experimental setup used and then explain how these are used to analyze the mobile web QoE. To objectively compare impact of server push on the web QoE and to measure the end-user perception we choose PLT as the metric [18], [19]. The PLT plays a vital role for both the end-user and the content provider: the end-user wants to view the content as early as possible, while the content provider has to increase the revenue by engaging the user longer. Our experimental setup consists of four components. The detailed discussion on how each component was selected follows.

Web Server: We used Jetty (release 9.4.11) [20], a Java-based open-source HTTP server. We choose Jetty from the list of tested implementations supporting HTTP/2 [21] since, we require the same server to support both HTTP/1.1 and HTTP/2 protocols. We activated additional SSL and HTTP/2 modules to add HTTPS and HTTP/2 support for the server. The HTTP/2 server push functionality was implemented using the experimental Servlet API provided by [22].

Mobile Client: We used a Dell Venue 8 Pro Tablet configured with Windows 8.1 operating system as the mobile client to carry out the experiments. We have used the Google Chrome browser's built-in network logger, in an incognito window with disable object caching, to capture the PLTs of the web pages.

Network Traffic Shaping: We conducted the experiments on two different network conditions: (i) on a mobile network trace-driven testbed and (ii) on a controlled wireless LAN with latency and loss variations. In the mobile network trace-driven testbed, our motive was to evaluate the web QoE using realistic mobile network dynamics. Accordingly, we used the traces from publicly available 3G/HSDPA dataset [23] and 4G/LTE bandwidth logs [24]. We chose traces collected using

bus and tram in the 4G and 3G mobile network conditions, respectively. These traces are useful for performing a good stress test for the two protocols, since the throughput is measured on short timescales (i.e., with 1 sec interval). We used Wondershaper [25], a simple traffic shaping script, to shape the bandwidth of client-server link. We throttled both the up-link and down-link capacity with the throughput values from the aforesaid trace files. In the controlled wireless LAN setup, we were able to use deterministic RTTs and PLR variations using NetEm utility [26] to simulate different mobile network conditions.

Limitations: We are aware that our experimental scenarios do not exactly replicate the reality and the limitations are listed below:

- We hosted all the websites on a single server, thus no *domain sharding* feature was supported.
- The websites were accessed using the IP address directly. Thus, the time taken to resolve the IP address is not included in our results.
- All the experiments were conducted on an isolated LAN and the mobile device was connected to a Wi-Fi router. There was no background traffic during the experiments.

Web Content: We have cloned the real websites and hosted locally for the controlled mobile web QoE assessment. We picked most popular websites from the list by Alexa [27], with high unique daily page views and high average time spent on a global scale. We mirrored 12 different websites as listed in Table I, across several popular Alexa categories. We used *HTTrack Website Copier* tool [28] to clone the websites along with their dependencies and the mirroring depth set to 1. Only the mobile version of the websites were cloned from the United States domains i.e., Amazon.com, not Amazon.in. The Table I also summarizes the resource statistics of the websites picked for the experiments. These values as reported in Table I (columns 6 to 13), i.e., the number of resources identified across each file type and their average file bundle size, play an important role because both the structure and resource bundle size of the website(s) significantly impact the PLT.

IV. RESULTS AND ANALYSIS

This section presents and analyzes the experimental results that demonstrate the implications of HTTP/2 server push feature on mobile web QoE. We categorize our results based on the various network conditions as: (i) web QoE under 4G/LTE network conditions, (ii) web QoE under 3G network conditions, (iii) impact of network latency on web QoE and (iv) impact of PLR on web QoE. For each scenario and for each website, the experiments were run 30 times and the average values are considered for the performance analysis. The experimental results have 95% confidence interval (CI).

A. Web QoE under 4G/LTE network

We emulate the network bandwidth using the 4G/LTE traces in a controlled manner (as discussed in Section III). Figure 2 compares the PLT for each website with both the protocols. We

TABLE I
STATISTICS OF WEBSITE RESOURCES USED FOR THE EXPERIMENTS

Sl #	Website Name	Category	Time spent per visitor (mm:ss)	Page views per visitor	Number of Resources				Average size (with SD) of the file bundle (in KB)			
					HTML	CSS	JS	Images	HTML	CSS	JS	Images
					10	1	38	54	229(682)	30(0)	123(179)	25(28)
1	CNN	News	04:02	2.11	9	2	93	125	34(91)	86(89)	23(61)	40(141)
2	Foxnews	News	05:33	2.57	7	32	53	50(120)	34(29)	59(81)	24(22)	
3	BBC	News	04:14	3.86	4	6	12	140	129(267)	70(88)	65(111)	16(17)
4	Amazon	Shopping	07:59	8.17	9	4	17	53	41(117)	37(27)	57(89)	77(334)
5	eBay	Shopping	09:21	7.07	4	2	14	52	179(255)	227(38)	71(68)	22(33)
6	iHerb	Shopping	10:48	7.39	13	7	24	22	108(261)	186(126)	96(187)	19(24)
7	Booking	Recreation	09:34	4.48	1	1	10	34	124(324)	331(357)	165(319)	31(18)
8	Emirates	Recreation	07:12	4.84	1	2	28	14	27(55)	228(0)	40(79)	43(36)
9	Hotels	Recreation	05:22	3.65	1	1	10	34	90(0)	53(0)	64(94)	7(5)
10	Cricbuzz	Sports	08:40	4.28	1	3	13	57	155(225)	364(503)	146(378)	56(56)
11	Nexusmods	Sports	12:17	10.60	1	5	11	24	73(0)	10(16)	103(77)	9(16)
12	Chess	Sports	08:53	4.87	1	1	1	1				

Note: Time spent per visitor and Page views per visitor are the estimated daily time spent on site (mm:ss) per visitor and unique page views per visitor on the site, respectively. These statistics are based on the trailing 3 months.

can see that all the websites records a significantly lower PLT with HTTP/2 compared to HTTP/1.1. It can also be seen from Figure 2 that the websites namely, CNN, Foxnews, Emirates and Nexusmods have recorded higher PLT when requested with HTTP/1.1 protocol. The prime reason for this higher PLT could be attributed to the larger number of large-sized JavaScript files associated with the web page(s). In case of HTTP/1.1, requesting and receiving the responses for these JavaScript files not only takes longer time (due to their file size) but also delays the subsequent object requests to be generated after parsing the scripts. This notably increases the PLT. However, in HTTP/2 these files and subsequent objects are prematurely pushed in an interleaved fashion using the multiplexing feature. This leads to a moderate PLT when HTTP/2 is used with the server push feature. Figure 3 illustrates the relative difference in the average PLT across the two protocols for different websites. We observed that HTTP/2 loads the websites 23% faster on an average, with 32% being the maximum improvement when compared to HTTP/1.1.

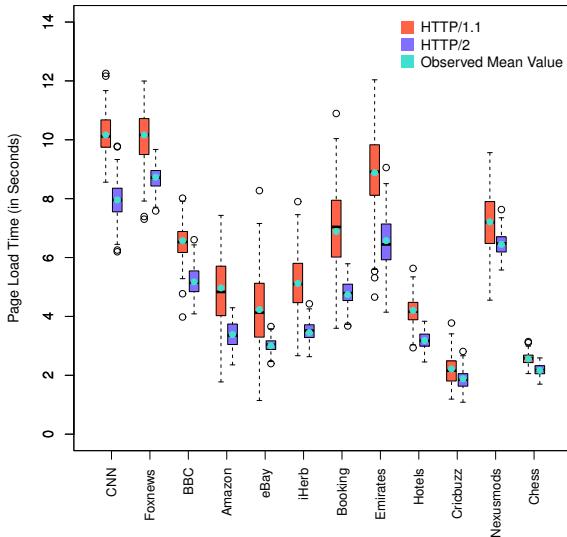


Fig. 2. Distribution of PLT observed for the websites when loaded with HTTP/2 and HTTP/1.1 under 4G/LTE trace-driven network condition.

Setting up a TCP connection requires three way handshake, which takes one and half RTT before sending the actual object request. This motivates us to investigate the number of TCP connections used to load a website using both the protocols. We observed that loading web page with HTTP/1.1 opens multiple TCP connections whereas, HTTP/2 transfer all the objects over a single TCP connection using the request and response multiplexing feature. Figure 4 depicts the number of TCP connections opened to load the web page with HTTP/1.1. As we know, the TCP user timeout control of 500 milliseconds (including both server processing delay and round-trip time), used for various TCP timeouts, prevents the client to reuse the TCP channel for sending additional requests in HTTP/1.1. Downloading a large-sized JavaScript file and generating the subsequent requests could take more than 500 milliseconds. This avoids reusing the existing TCP connection(s) in HTTP/1.1 and leads to opening multiple TCP connections. As seen from Figure 4, similar trend was observed for all the websites with large-sized JS files as their dependencies.

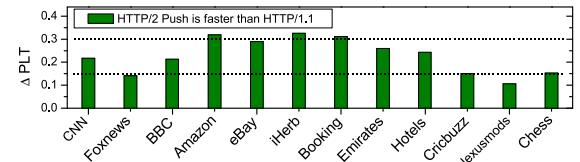


Fig. 3. The relative difference in the PLT (Δ PLT) under 4G/LTE trace-driven network condition.

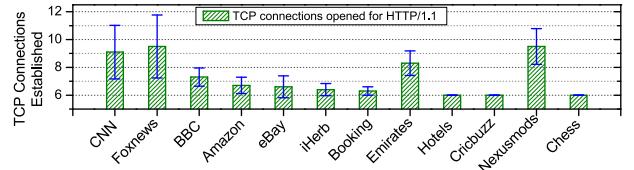


Fig. 4. Number of TCP connections established when web page loaded with HTTP/1.1 under 4G/LTE trace-driven network conditions.

B. Web QoE assessment under 3G network

Next, we investigated the web QoE under 3G network condition. The PLT for all the websites when loaded with both HTTP/1.1 and HTTP/2 under 3G network condition are shown in Figure 5. We observe that PLT with HTTP/2 is significantly lower as compared with HTTP/1.1. From Figure 6, it is also evident that HTTP/2 relatively loads the web pages 24% faster on an average - with 40% being the maximum improvement - compared to HTTP/1.1. Furthermore, we also observed that websites with fewer objects has higher relative difference in PLT (Δ PLT) when loaded with HTTP/2. On the contrary, all the other websites have moderate values of Δ PLT. A trend similar to the one with 4G network condition was observed as well for the TCP connection overhead as reported in Figure 7. The significant increase in the number of TCP connections (for all the websites with HTTP/1.1) as compared to the 4G network condition can be ascribed to the lower available bandwidth in the 3G network.

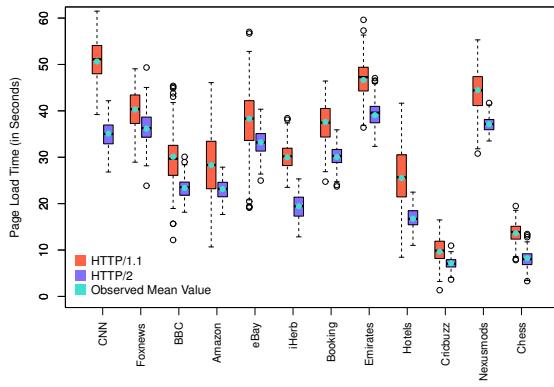


Fig. 5. Distribution of PLT observed for the websites when loaded with both HTTP/2 and HTTP/1.1 under 3G trace-driven network condition.

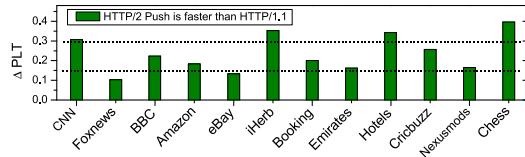


Fig. 6. The relative difference in the observed PLT (Δ PLT) under 3G trace-driven network condition.

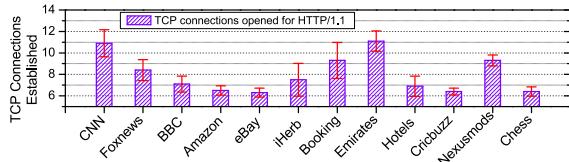


Fig. 7. The TCP connection overhead when web page loaded with HTTP/1.1 under 3G trace-driven network condition.

C. Impact of network latency on web QoE

To assess the impact of the network latency on web QoE, we emulate the link with four RTTs; 0 ms, 100 ms, 200 ms

and 300 ms. We also restricted the available link bandwidth to 1 Mbps (representing access bandwidth in developing countries). For this experiment, we chose four websites, CNN, Foxnews, Amazon and Emirates from the set listed in Table I. The PLT variation for these websites observed with the network latency variations are shown in Figure 8. We observed that the PLT for all the websites increases with increasing RTT values both for HTTP/1.1 and HTTP/2. However, the PLT observed with HTTP/2 is always lower than the PLT with HTTP/1.1. From Figure 8, we conclude that HTTP/2 with server push feature gives better QoE even in networks with higher RTTs.

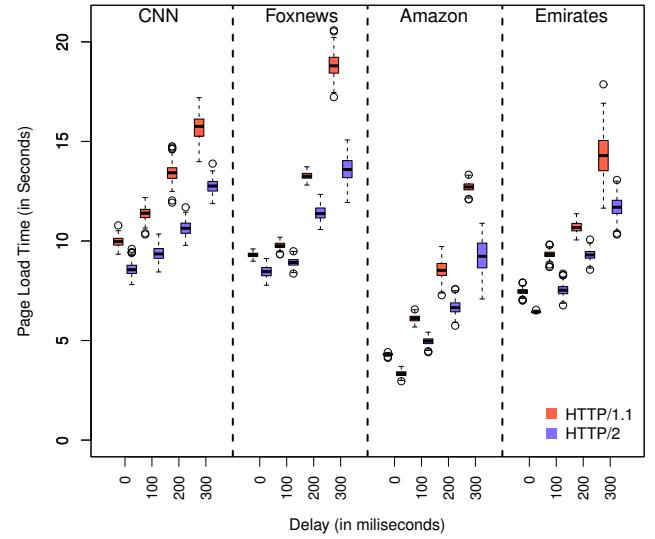


Fig. 8. Distribution of PLT observed for both HTTP/2 and HTTP/1.1 under network latency variations.

D. Impact of PLR on web QoE

The packet loss ratio (PLR) is also an important aspect of mobile networks. Accordingly, we also conducted experiments varying the PLR while using fixed network latency of 100ms and the link bandwidth set to 1 Mbps. We used three different PLR values; 0%, 1%, and 2%. The relative difference in PLT for the websites under the PLR variations is shown in Figure 9. Results reveal that HTTP/1.1 outperforms HTTP/2 and loads the web pages faster when the network PLR is 2% or more (see Figure 9). The reversing trend when the PLR is higher because HTTP/2 uses a single TCP connection to transfer the objects from the server. In a network with higher PLR, all the HTTP/2 streams sharing a single TCP connection are more susceptible to packet loss and require retransmission(s). This will contribute to an increase in the PLT. On the contrary, HTTP/1.1 opens up multiple TCP connections to address the packet loss problem. This is also corroborated by the results shown in Figure 10, where we observe that the number of TCP connections with HTTP/1.1 increases as the PLR increases.

V. CONCLUSION

In this paper, we experimentally studied the implications of HTTP/2 on mobile web QoE. The study was conducted to

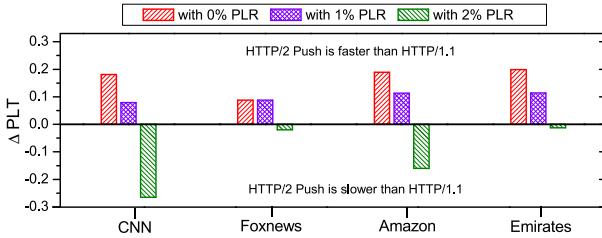


Fig. 9. The relative difference in PLT (Δ PLT) for placket loss ratio variations.

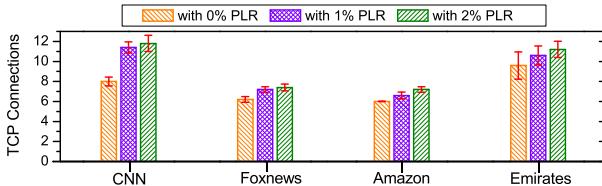


Fig. 10. Impact of packet loss ratio variation on the number of TCP connections established with HTTP/1.1.

understand if the HTTP/2 server push feature really improves the web QoE in mobile devices, in terms of the popular objective web QoE metric page load time (PLT). We selected 12 popular websites from Alexa ranking and mirrored them for conducting the controlled experiments. By conducting the experiments in a trace-driven testbed with both 4G/LTE and 3G network traces, we observed that HTTP/2 outperforms HTTP/1.1 in terms of the PLT. Our experimental results shows that, on an average there is 24% lower PLT with HTTP/2. We also observed a significant TCP connection overhead in HTTP/1.1 under both the network conditions. This not only increases the load on the server (in terms of maintaining the state information) but also consumes more energy on the mobile devices. We also investigated the impact of network latency and PLR on mobile web QoE. Our results reveal that HTTP/2 performs quite well even in networks with higher round trip times. However, we identified the condition(s) under which the HTTP/2 is not very good. Our results demonstrate that HTTP/1.1 gives lower PLT than HTTP/2 as the PLR increases. Our future research will continue to focus on conducting the mobile web QoE experiments to compare the effects of alternate network configurations and investigating the impact of HTTP/2 push feature with an emphasis on the relation to the web page characteristics.

REFERENCES

- [1] Cisco Systems, Inc., "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016 - 2021," Cisco White Paper, March 2017.
- [2] *Vocabulary for performance and quality of service. Amendment 2: New definitions for inclusion in Recommendation ITU-T P.10/G.100*, ITU-T Recommendation P.10/G.100, July 2008.
- [3] J. C. Mogul, "The Case for Persistent-connection HTTP," *SIGCOMM Comput. Commun. Rev.*, vol. 25, no. 4, pp. 299–313, Oct 1995.
- [4] M. Butkiewicz, H. V. Madhyastha, and V. Sekar, "Understanding Website Complexity: Measurements, Metrics, and Implications," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC '11)*. New York, NY, USA: ACM, 2011, pp. 313–328.
- [5] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, and H. Yan, "Modeling Web Quality-of-experience on Cellular Networks," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking (MobiCom '14)*. New York, NY, USA: ACM, 2014, pp. 213–224.
- [6] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the Quality of Experience of Web Users," *SIGCOMM Comput. Commun. Rev.*, vol. 46, no. 4, pp. 8–13, Dec 2016.
- [7] *The http archive*, [Online]. Available: <http://httparchive.org/trends.php>.
- [8] M. Belshe, R. Peon, and E. M. Thomson, "Hypertext Transfer Protocol Version 2 (HTTP/2)," May 2015, [Online]. Available: <https://tools.ietf.org/html/rfc7540>.
- [9] "SPDY Protocol - Draft 3.1," [Online]. Available: <http://www.chromium.org/spdy/spdy-protocol/spdy-protocol-draft3-1>.
- [10] M. Varvello, K. Schomp, D. Naylor, J. Blackburn, A. Finamore, and K. Papagiannaki, "Is the Web HTTP/2 Yet?" in *Proceedings of the 17th International Conference on Passive and Active Measurement (PAM 2016)*. Heraklion, Greece, March 2016, pp. 218 – 232.
- [11] L. M. Bach, B. Mihaljevic, and A. Radovan, "Exploring HTTP/2 advantages and performance analysis using Java 9," in *Proceedings of the 40th Intl. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2017, pp. 1522 – 1527.
- [12] D. Strohmeier, S. Egger, A. Raake, T. Hörfeld, and R. Schatz, *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer, 2014, ch. Web Browsing, pp. 329 – 338.
- [13] D. Guse, S. Schuck, O. Hohlfeld, A. Raake, and S. Müller, "Subjective quality of webpage loading: The impact of delayed and missing elements on quality ratings and task completion time," in *Proceedings of the Seventh International Workshop on Quality of Multimedia Experience (QoMEX '15)*, May 2015, pp. 1 – 6.
- [14] C. Kelton, J. Ryoo, A. Balasubramanian, and S. R. Das, "Improving User Perceived Page Load Times Using Gaze," in *Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation (NSDI '17)*, Boston, MA, USA, March 2017.
- [15] S. A. Chowdhury, V. Sapra, and A. Hindle, "Is HTTP/2 More Energy Efficient Than HTTP/1.1 for Mobile Users?" *PeerJ PrePrints*, vol. 3, no. e1280v1, August 2015.
- [16] V. Sapra and A. Hindle, "Web Servers Energy Efficiency Under HTTP/2," *PeerJ PrePrints*, vol. 4, no. e2027v1, May 2016.
- [17] T. Zimmermann, B. Wolters, and O. Hohlfeld, "A QoE Perspective on HTTP/2 Server Push," in *Proceedings of the Workshop on QoE-based Analysis and Management of Data Communication Networks (Internet QoE '17)*. New York, NY, USA: ACM, August 2017, pp. 1–6.
- [18] U. Goel, M. Steiner, M. P. Wittie, M. Flack, and S. Ludin, "Poster:HTTP/2 Performance in Cellular Networks," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. New York, NY, USA: ACM, 2016, pp. 433–434.
- [19] H. de Saxe, I. Oprescu, and Y. Chen, "Is HTTP/2 really faster than HTTP/1.1?" in *Proceedings of the IEEE Conference on Computer Communications Workshops*, April 2015, pp. 293–299.
- [20] "Eclipse Jetty," [Online]. Available: <http://www.eclipse.org/jetty/>.
- [21] M. Wangen, "Known implementations of HTTP/2," 2018, [Online]. Available: <https://github.com/http2/http2-spec/wiki/Implementations>.
- [22] Webtide, *HTTP/2 Push with experimental Servlet API*, August 2014, [Online]. Available: <https://webtide.com/http2-push-with-experimental-servlet-api/> (Last accessed: September 2018).
- [23] H. Riiser, T. Endestad, P. Vigmostad, C. Griwodz, and P. Halvorsen, "Video Streaming Using a Location-based Bandwidth-lookup Service for Bitrate Planning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, no. 3, pp. 24:1–24:19, Aug 2012.
- [24] J. van der Hooft, S. Petrangeli, T. Wauters, R. Huysegems, P. R. Alface, T. Bostoen, and F. De Turck, "HTTP/2-Based Adaptive Streaming of HEVC Video Over 4G/LTE Networks," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2177–2180, 2016.
- [25] Wondershaper - Traffic Shaping Script, [Online]. Available: <https://packages.debian.org/unstable/net/wondershaper>.
- [26] S. Hemminger, "Network emulation with NetEm," in *Linux Conf Au*, April 2005, pp. 18–23.
- [27] Alexa, *Alexa Top 500 sites on the web*, [Online]. Available: <http://www.alexa.com/topsites> (Last accessed: September 2018).
- [28] "HTTrack Website Copier - Version 3.49-2," May 2017, [Online]. Available: <https://www.httrack.com/>.

A Graph Based Clustering and Preconditioning of V-MIMO Wireless Sensor Networks

Rakesh Mundlamuri, Thangapandian B, Vijay K Chakka and Srikanth Goli

Department of Electrical Engineering

Shiv Nadar University

NH - 91, Tehsil Dadri, Gautam Buddha Nagar, Uttar Pradesh - 201314, India

Email: rm845@snu.edu.in, tb693@snu.edu.in, vk484@snu.edu.in, gs499@snu.edu.in

Abstract—This paper presents a graph based methodology for increasing the channel capacity of Virtual-Multiple Input Multiple Output (V-MIMO) defined over a Wireless Sensor Network (WSN). A fully connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{W})$ is defined for a WSN. Then, we propose a new clustering algorithm based on the Fiedler vector of the graph \mathcal{G} which divides the sensor nodes \mathcal{V} into two clusters (transmitting and receiving antennas). The links between these two clusters results in V-MIMO network. Next, a Modified Maximum Spanning Tree Search algorithm (MMASTS) is proposed on V-MIMO to enhance the average channel capacity. Simulation performance of average channel capacity and uncoded Bit Error Rate (BER) are plotted using different precoding techniques like Zero Forcing (ZF) and Minimum Mean Square Error (MMSE). These are also used for comparing the performance of proposed Fiedler vector based clustering with k -means clustering.

Index Terms—WSN , Fiedler Vector, V-MIMO, MMASTS Algorithm, ZF and MMSE.

I. INTRODUCTION

Wireless Sensor Networks (WSN) are a group of spatially dispersed, dedicated sensors for monitoring and recording the physical conditions of the environment. These sensors are constrained by limited processing capability, physical size, limited battery and single antenna [1], [2]. The data extraction from these sensors is critical in many practical applications of WSN. To increase the data rate among these sensors, V-MIMO concept was first introduced by Mischa et.al in [3]. In this method, the sensor head of WSN divides the sensor nodes into two clusters each having a cluster head by providing virtual links among them. This results in the V-MIMO structure.

In [4], [5], [6] Jayaweera et.al introduced cooperative communication based V-MIMO among the sensors in WSN for low energy consumption. LEACH algorithm and Multi-hop MIMO LEACH scheme [8], [9] are some of the widely used clustering algorithms in WSN which primarily focus on optimizing energy consumption. Liang Jing et.al in [7] extended virtual MIMO concept to increase the channel capacity by proposing an MASTS algorithm with *pre-defined two sensor clusters*. In this paper, one of the graph based clustering methods called Fiedler vector based clustering is proposed for WSN. It gives the flexibility to choose the nodes and enhance the channel capacity in the case of fast fading channels. This clustering algorithm is compared with one of the standard clustering

algorithm called k-means clustering algorithm [10]. In this clustering method, Fiedler vector [11], [12] of the graph is used to decompose the WSN into two clusters as shown in Fig.1. This formation of V-MIMO network may result in a singular channel matrix similar to k-means, which degrades the average channel capacity. To overcome this limitation, a Modified MASTS (MMASTS) algorithm is proposed to improve the rank as well as condition number of the V-MIMO channel matrix. This improves the average channel capacity.

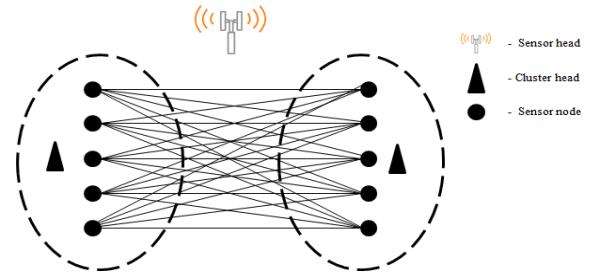


Fig. 1. Virtual MIMO Model

The rest of the paper is discussed as follows: Section II deals with the graph based clustering and low-rank V-MIMO system model. Section III presents graph based MMASTS algorithm. Section IV describes the precoder based average channel capacity. Section V discussed the simulation results to verify the performance of MMASTS using ZF and MMSE precoders. Finally, conclusions are drawn in section VI.

Notations : $(\cdot)^H$ and $(\cdot)^{-1}$ are the Hermitian (conjugate and transpose) and the inverse of a matrix. \mathbf{I}_N represents the Identity matrix of size N . $\mathcal{R}(\cdot)$ denotes the rank of a matrix.

II. GRAPH BASED CLUSTERING MODEL

Assume a sensor head having a processing capability to control all the nodes of WSN and two cluster heads having the processing capability to control sensor nodes in their respective clusters as shown in Fig. 1. Let us also consider a complete/connected graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{W})$ [13], where nodes (\mathcal{V}) as sensors, edges (\mathcal{E}) as links between them and fading real channel coefficients as weights (\mathcal{W}). A weighted Adjacency matrix \mathbf{A} and un-normalized Laplacian matrix \mathbf{L} represents the graph \mathcal{G} of WSN.

A. Formation of Two Clusters

The un-normalized Laplacian matrix \mathbf{L} of a fully connected graph is obtained by,

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \quad (1)$$

where \mathbf{D} is the degree matrix [14] of \mathcal{G} . Eigenvector of the least non-zero eigenvalue of \mathbf{L} is known as Fiedler vector. The sensor head which controls all the nodes decomposes the WSN into two clusters having equal number of nodes using this Fiedler vector.

Steps to implement the Graph clustering algorithm:

1. Find all eigenvalues $(\lambda_1, \lambda_2, \dots, \lambda_n)$ and their corresponding eigenvectors (u_1, u_2, \dots, u_n) of the Laplacian matrix defined above.
2. Sort the obtained eigenvalues in ascending order i.e., $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ and its corresponding eigenvectors as u_1, u_2, \dots, u_n .
3. Determine the median of the u_1 vector elements.
4. Clusters are formed based on the above computed median as a threshold i.e., nodes with indices greater than the threshold are segregated into one cluster and the rest into another.
5. Nodes are re-indexed in each cluster.

The rearranged adjacency matrix according to the new indices is given below:

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{G}_1 & \mathbf{H}_1 \\ \mathbf{H}_2 & \mathbf{G}_2 \end{bmatrix}, \quad (2)$$

where \mathbf{G}_1 and \mathbf{G}_2 are the transmit and receive correlation matrices (intracluster channels). \mathbf{H}_1 and \mathbf{H}_2 are the uplink and downlink channel matrices (intercluster channels) respectively. For Time Division Duplexing (TDD) and uncorrelated intracluster channels the adjacency matrix is modified by $\mathbf{H}_1 = \mathbf{H}$, $\mathbf{H}_2 = \mathbf{H}^T$ and $\mathbf{G}_1 = \mathbf{G}_2 = \mathbf{0}$. For Frequency Division Duplexing (FDD) and uncorrelated intracluster channels $\mathbf{H}_1, \mathbf{H}_2$ are independent. Throughout the paper, we assume TDD with uncorrelated intracluster channels. Thus the new adjacency matrix is reduced to

$$\mathbf{A}' = \begin{bmatrix} \mathbf{0} & \mathbf{H} \\ \mathbf{H}^T & \mathbf{0} \end{bmatrix}. \quad (3)$$

This new adjacency matrix represents a bipartite structure as shown in Fig.1. This also represent as V-MIMO network. Since the \mathbf{H} in (3) is based on virtual links, which may result in singular and ill-conditioned matrices.

B. Low Rank V-MIMO System Model

The above clustered sensors forming a simplex mode is defined as,

$$\mathbf{y} = \mathbf{Hx} + \mathbf{n}, \quad (4)$$

where, $\mathbf{y} \in \mathbb{C}^{M_r \times 1}$ be the received signal vector, $\mathbf{x} \in \mathbb{C}^{M_t \times 1}$ is the transmitted signal vector, noise vector $\mathbf{n} \in \mathbb{C}^{M_r \times 1}$ denotes the Additive White Gaussian Noise with $\mathcal{N}(0, \sigma^2)$

and $\mathbf{H} \in \mathbb{R}^{M_r \times M_t}$ is the channel matrix. Each element in \mathbf{H} is *iid* real Gaussian random variable. $\mathcal{N}(0, 1)$ representing a flat fading model. M_t and M_r are the number of transmitting and receiving antennas respectively.

III. GRAPH BASED MODIFIED MASTS ALGORITHM

In this section, MASTS algorithm presented in [7], is modified to improve the rank of the channel matrix \mathbf{H} .

Modified MASTS algorithm:

1. Consider the \mathbf{H} matrix from (3). Let us consider an example of a 3x3 \mathbf{H} matrix,

$$\mathbf{H} = \begin{bmatrix} h_{11} = 0.4541 & h_{12} = 0.0637 & h_{13} = 0.5549 \\ h_{21} = 0.2009 & h_{22} = 0.3301 & h_{23} = 0.9422 \\ h_{31} = 0.5148 & h_{32} = 0.3717 & h_{33} = 0.5529 \end{bmatrix}. \quad (5)$$

2. Convert the \mathbf{H} matrix into a vector form as given below.

$$\mathbf{H} = \begin{bmatrix} h_{11} = 0.4541 \\ h_{12} = 0.0637 \\ h_{13} = 0.5549 \\ h_{21} = 0.2009 \\ h_{22} = 0.3301 \\ h_{23} = 0.9422 \\ h_{31} = 0.5148 \\ h_{32} = 0.3717 \\ h_{33} = 0.5529 \end{bmatrix}. \quad (6)$$

3. Sort the elements of the vector in decreasing order without losing the indices as given below.

$$\mathbf{H} = \begin{bmatrix} h_{23} = 0.9422 \\ h_{13} = 0.5549 \\ h_{33} = 0.5529 \\ h_{31} = 0.5148 \\ h_{11} = 0.4541 \\ h_{32} = 0.3717 \\ h_{22} = 0.3301 \\ h_{21} = 0.2009 \\ h_{12} = 0.0637 \end{bmatrix}. \quad (7)$$

4. Select the first 3 elements from the step 3 and form a subgraph having edges representing indices.

$$\mathbf{H} = \begin{bmatrix} h_{23} = 0.9422 \\ h_{13} = 0.5549 \\ h_{33} = 0.5529 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (8)$$

5. Expand the subgraph created in step 4 by adding a next element from the step 3.

$$\mathbf{H} = \begin{bmatrix} h_{23} = 0.9422 \\ h_{13} = 0.5549 \\ h_{33} = 0.5529 \\ h_{31} = 0.5148 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (9)$$

6. Check whether the subgraph is forming a cyclic graph. If it forms a cyclic graph, then remove the last element and add the next from the step 3.

$$\mathbf{H} = \begin{bmatrix} h_{23} = 0.9422 \\ h_{13} = 0.5549 \\ h_{33} = 0.5529 \\ h_{31} = 0.5148 \\ 0 \\ h_{32} = 0.3717 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \quad (10)$$

7. Repeat step 5 and step 6 until the expanded subgraph contains the edge number equal to $(M_t + M_r - 1)$. This subgraph results in the maximum spanning tree [15] forming a new matrix $\bar{\mathbf{H}}$.

$$\bar{\mathbf{H}} = \begin{bmatrix} 0 & 0 & h_{13} = 0.5549 \\ 0 & 0 & h_{23} = 0.9422 \\ h_{31} = 0.5148 & h_{32} = 0.3717 & h_{33} = 0.5529 \end{bmatrix}. \quad (11)$$

8. Check whether the above-formed matrix $\bar{\mathbf{H}}$ is a singular or a non-singular matrix.

- a. If step-8 results into a non-singular matrix then exit.
- b. If step-8 results in a singular matrix then add the maximum weighted edge picked from the excluded edge pool forming a new subgraph with increment in number of edges defined in step 7. Repeat step 8.

$$\bar{\mathbf{H}} = \begin{bmatrix} h_{11} = 0.4541 & 0 & h_{13} = 0.5549 \\ 0 & 0 & h_{23} = 0.9422 \\ h_{31} = 0.5148 & h_{32} = 0.3717 & h_{33} = 0.5529 \end{bmatrix}. \quad (12)$$

This process results in a new channel matrix $\bar{\mathbf{H}}$ with improved rank of the channel matrix \mathbf{H} .

IV. PRECODER AND CHANNEL CAPACITY

Now, the new channel matrix obtained using MMASTS algorithm $\bar{\mathbf{H}}$ is used for precoding. The data vector $\mathbf{x} = \mathbf{Is}$ defined in equation (4) is transformed using the precoder matrix \mathbf{F} as $\hat{\mathbf{x}} = \mathbf{Fs}$. Here s is the desired information symbols for the M_r nodes. Based on the perfect Channel State

Information ($CSI \equiv \bar{\mathbf{H}}$) at the transmitter, the ZF [16] and MMSE [17] precoder $\mathbf{F} \in \mathbb{R}^{M_t \times M_r}$ are as follows,

$$\mathbf{F} = \begin{cases} \bar{\mathbf{H}}^H (\bar{\mathbf{H}} \bar{\mathbf{H}}^H)^{-1}, & \text{ZF} \\ \bar{\mathbf{H}}^H (\bar{\mathbf{H}} (\bar{\mathbf{H}}^H)^H + \sigma^2 \mathbf{I})^{-1}, & \text{MMSE} \end{cases}. \quad (13)$$

MMSE requires the assumption of additive white Gaussian noise variance at the transmitter itself, whereas ZF does not require noise variance. To show the performance of the proposed MMASTS algorithm on the V-MIMO network, we use the channel capacity and BER as a performance metrics.

Channel Capacity:

Channel capacity with equal power allocation $(\frac{P}{M_t})$ is represented as [18] [19],

$$C = \sum_{i=1}^{\mathcal{R}(\bar{\mathbf{H}})} \log_2 \left(1 + \frac{\bar{\lambda}_i SNR_i}{M_t} \right), \quad (14)$$

where, $SNR_i = \frac{P}{\sigma^2}$ is the signal to noise ratio at the i^{th} receiver node. P is total power given, $\bar{\lambda}_i$ is the i^{th} eigenvalue of $\bar{\mathbf{H}} \bar{\mathbf{H}}^H (\bar{\mathbf{H}} = \bar{\mathbf{H}} \mathbf{F})$ and σ^2 is the noise variance.

V. SIMULATION RESULTS

In this section, we show the simulation results conducted using 5000 channel realizations. A 16-QAM modulation scheme is used for representing data for different MIMO configurations like 8×8 , 16×16 and 32×32 . The performance of MASTS algorithm [7] and the proposed MMASTS algorithm are tabulated in TABLE I. This shows the percentage of non-singular matrices resulted from the above channel realizations and MIMO configurations. From TABLE I, MASTS algorithm results in zero percentage of non-singular matrices for MIMO configurations 32×32 and above. Whereas, proposed MMASTS algorithm results in 27.22% of non-singular matrices for 32×32 . As size of MIMO configuration increases the percentage of non-singular matrices resulting from both MASTS and MMASTS are falling.

TABLE I
PERCENTAGE NON-SINGULAR MATRICES

$M_r \times M_t$	MASTS%	MMASTS%
8×8	8.64	91.3
16×16	0.12	61.78
32×32	0	27.22

The proposed Graph clustering method is compared with the k-means clustering method without applying any MASTS and MMASTS algorithms. Fig. 3 and Fig. 6 shows these results. Fig. 3 shows the better performance of proposed Fiedler vector based clustering over k-means clustering in terms of average channel capacity using a full rank channel matrix realizations. Whereas, deterioration of performance in the case of low rank channel matrix realizations is observed clearly from Fig. 6. The significant improvement over Fig. 6 is shown in Fig. 4 and Fig. 7 using Fiedler vector based clustering with inclusion of MASTS (or) MMASTS algorithm and ZF (or)

MMSE precoders. The reason for this improvement is evident from Table II and Fig. 2. Fig. 5 shows the channel capacity performance comparison of MMSE and ZF precoders along with Fiedler vector clustering and MMASTS algorithm. Fig. 8 shows the uncoded BER performance of lowrank V-FULL MIMO, V-MMASTS MIMO and V-MASTS MIMO using ZF and MMSE precoders.

Let $\bar{\lambda}_{max}$ and $\bar{\lambda}_{min}$ be the maximum and minimum eigenvalue of $\hat{\mathbf{H}}\hat{\mathbf{H}}^H$ respectively. Therefore, the count of the channel matrices which are less than the desired condition number ($\frac{\bar{\lambda}_{max}}{\bar{\lambda}_{min}}$) with Modified MASTS algorithm along with ZF, MMSE precoders for 5000 channel realizations are tabulated as follows.

TABLE II
MATRIX COUNT FOR DIFFERENT CONDITION NUMBERS

Condition number \ Precoder	= 1	≤ 10	≤ 30	≤ 50
ZF	91	4926	4926	4926
MMSE	90	3210	3783	3942

From Table II, we can see that the modified MASTS algorithm with ZF precoder has better channel condition than with MMSE precoder. This also evident from the CDF plot of condition numbers depicted in Fig. 2.

In general, MMSE precoder is used to reduce the condition number of ill-conditioned matrices. This is the traditionally reason, where MMSE based precoder always outperforms the ZF precoder. In this paper, from Fig. 5. showing an opposite trend. The reason for this opposite trend is that our proposed modified MASTS algorithm is improving the condition number of ill-conditioned matrices. Table-II and Fig. 2, clearly shows that the modified MASTS followed by ZF precoder has better condition number than the MMSE precoder, which results in opposite action of the traditional trend.

VI. CONCLUSION

In this paper, we proposed a Fiedler vector based cluster formation for creation of V-MIMO in WSN. The channel capacity and BER performances of V-MIMO is improved significantly by using MASTS and proposed MMASTS algorithms along with Fiedler vector based clustering.

ACKNOWLEDGMENT

We would like to show our gratitude to Dr. Satyanarayana Reddy, we are also thankful to Mr. Basheeruddin Shah Shaik and Mr. Venugopalachary for their comments on preparing the manuscript.

REFERENCES

- [1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, Aug 2002.
- [2] A. D. Coso, U. Spagnolini, and C. Ibars, "Cooperative distributed MIMO channels in wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 2, pp. 402–414, February 2007.
- [3] M. Dohler *et al.*, "Virtual antenna arrays," Ph.D. dissertation, University of London, 2004.
- [4] S. K. Jayaweera, "Virtual MIMO-based cooperative communication for energy-constrained wireless sensor networks," *IEEE Transactions on Wireless Communications*, vol. 5, no. 5, pp. 984–989, May 2006.
- [5] ———, "Energy efficient virtual MIMO-based cooperative communications for wireless sensor networks," in *Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing, 2005.*, Jan 2005, pp. 1–6.
- [6] ———, "An energy-efficient virtual MIMO architecture based on V-BLAST processing for distributed wireless sensor networks," in *2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, 2004. IEEE SECON 2004*, Oct 2004, pp. 299–308.
- [7] J. Liang and Q. Liang, "A graph theoretical algorithm for virtual MIMO channel selection in wireless sensor networks," in *MILCOM 2008 - 2008 IEEE Military Communications Conference*, Nov 2008, pp. 1–6.
- [8] Y. Yuan, M. Chen, and T. Kwon, "A novel cluster-based cooperative MIMO scheme for multi-hop wireless sensor networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2006, no. 1, p. 072493, 2006.
- [9] K. G. Panda, D. Agrawal, and A. Hossain, "Virtual MIMO in wireless sensor network-a survey," in *Green Engineering and Technologies (ICGET), 2016 Online International Conference on*. IEEE, 2016, pp. 1–4.
- [10] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 881–892, 2002.
- [11] M. Fiedler, "Laplacian of graphs and algebraic connectivity," *Banach Center Publications*, vol. 25, no. 1, pp. 57–70, 1989.
- [12] D. V. D. Ville, R. Demesmaeker, and M. G. Preti, "When Slepian Meets Fiedler: Putting a Focus on the Graph Spectrum," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1001–1004, July 2017.
- [13] D. B. West *et al.*, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.
- [14] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.
- [15] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.
- [16] N. K. D. Venkategowda, H. Lee, and I. Lee, "Data Precoding and Energy Transmission for Parameter Estimation in MIMO Wireless Powered Sensor Networks," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Sept 2017, pp. 1–5.
- [17] Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, *MIMO-OFDM wireless communications with MATLAB*. John Wiley & Sons, 2010.
- [18] A. Goldsmith, *Wireless communications*. Cambridge university press, 2005.
- [19] A. F. Molisch, M. Z. Win, Y.-S. Choi, and J. H. Winters, "Capacity of MIMO systems with antenna selection," *IEEE Transactions on Wireless Communications*, vol. 4, no. 4, pp. 1759–1772, July 2005.

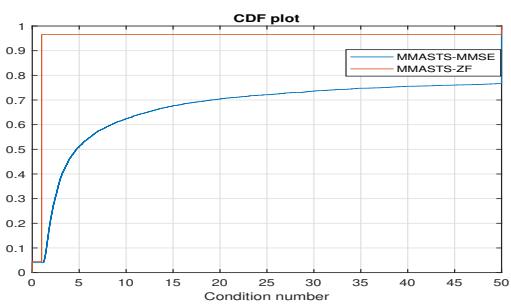


Fig. 2. CDF plot of the condition number given in TABLE II

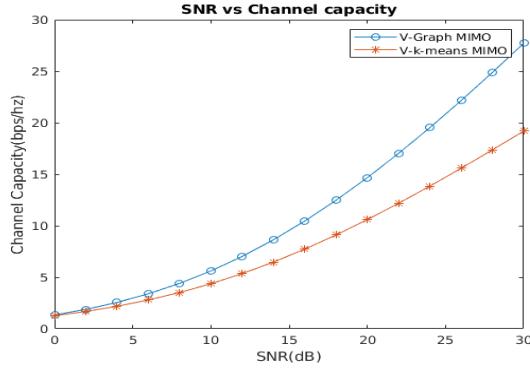


Fig. 3. SNR vs channel capacity of full rank channel matrices comparing Graph clustering and k-means

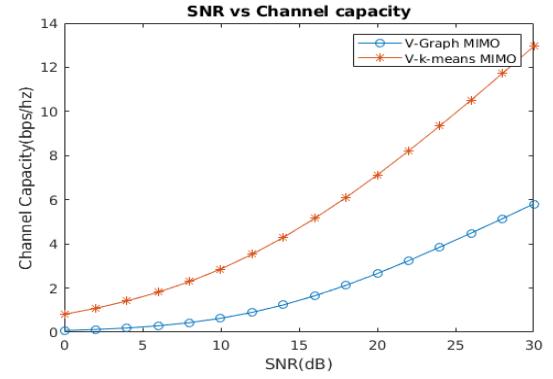


Fig. 6. SNR vs channel capacity of low rank channel matrices comparing Graph clustering and k-means

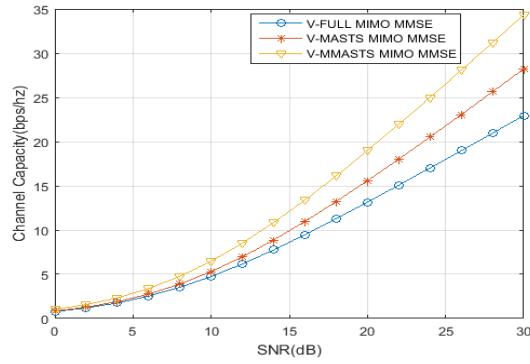


Fig. 4. SNR vs Channel capacity for low rank channel matrix with MMSE and $\sigma^2 = 0.01$

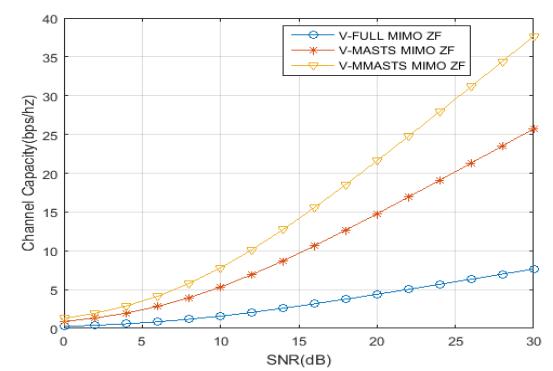


Fig. 7. SNR vs Channel capacity for low rank channel matrix with ZF

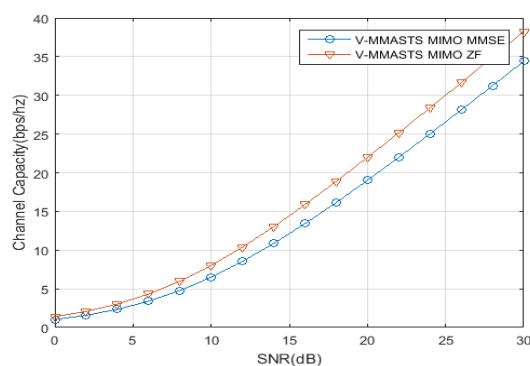


Fig. 5. SNR vs Channel capacity for low-rank channel matrix with ZF and MMSE - $\sigma^2 = 0.01$

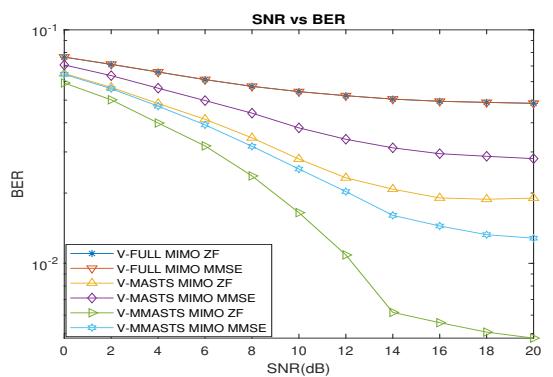


Fig. 8. SNR vs BER for lowrank channel matrix with ZF and MMSE - $\sigma^2 = 0.01$

Analysis of Mid-Haul Characteristics for LTE-NR Multi-Connectivity in Heterogeneous Cloud RAN

Ramakrishnan S, Subrat Kar, S Dharmaraja
Bharti School of Telecommunication Technology and Management
Indian Institute of Technology Delhi
New Delhi 110016, India

Email: ramakrishnan.s@ieee.org, subrat@ee.iitd.ac.in, dharmar@maths.iitd.ac.in

Abstract—To address the capacity requirements resulting from huge growth in mobile data traffic, the mobile network operators (MNOs) are deploying heterogeneous Cloud based Radio Access Network (C-RAN) with a mix of Base stations *viz.* 4G Macro Base stations (offering better coverage) and 5G Small cells (offering better radio capacities). With LTE-NR Multi-Connectivity feature in 5G network, the User Equipments (UEs) can connect with both 4G and 5G simultaneously to take advantage of better coverage and capacity, so that its QoS is met appropriately. To maintain such simultaneous connections (with UE) across different BTS, the network must support stringent high bandwidth and low-latency links (mid-haul links) between the Base station nodes. In a C-RAN environment, the inter Base station mid-haul links can be viewed as inter Virtual Machine (VM) communication link.

In this paper, we analyze the characteristics of the mid-haul link for LTE-NR Multi-Connectivity feature (specifically E-UTRAN-NR Dual Connectivity EN-DC configuration) in a heterogeneous C-RAN deployment. We simulate the C-RAN scenario using NS3 network simulator with heterogeneous mix of 4G and 5G-NR Base stations with appropriate cloud architecture split. We use the BTS power model to derive the Computational Requirement (CR) of the Base station for vBBU (virtual Baseband Unit) placement algorithm in C-RAN. Through simulation, we review the bandwidth requirement of the mid-haul link and assess variation in the end-to-end delay when latency of mid-haul link is varied. From available cloud infrastructure hardware bench-mark results, we observe that the latency of inter VM communication links, varies depending on whether the two communicating VMs are on same or different Cloud servers. Thus, we propose "Neighbour Association-aware Placement" (NAP) algorithm for placement of the vBBUs in the same Cloud server and assess the benefits in the case of EN-DC configuration.

I. INTRODUCTION

In the recent past, we have witnessed a tremendous growth in mobile data consumption. As per [1], the mobile data traffic has grown 18-folds globally in the last 5 years. It has increased by more than 63% in the past year and is expected to increase 7-folds in next 5 years. In a developing telecommunication market like India, we observe significant mobile data growth of around $\sim 144\%$ in a single year [2].

The requirement for such high capacity, is addressed by the mobile network operators (MNOs) through network densification, by deploying a mix of Base stations *viz.* 4G Macro Base stations (operating at lower frequency, offering better coverage) and 5G Small cells (operating at higher frequency,

offering better radio capacities). This is highlighted in Table I.

TABLE I
COVERAGE AND CAPACITY LAYERS

Spectrum Range	Coverage	Bandwidth	Layer type	Duplexing
< 2Ghz	Higher	Lower	Coverage	FDD
> 2GHz	Lower	Higher	Capacity	TDD

Thus, the MNOs need to develop a strategy to position each User Equipment (UE) to specific layer to make sure the Quality of Service (QoS) criterion of UE (in terms of coverage and capacity) is met. This is part of the MNO's overall UE camping strategy.

Traditionally, the MNO's UE camping strategy have assumed UEs that can connect to at most one spectrum layer of the network. The 5th Generation (5G) network introduces more flexible Multi-Connectivity feature, that allows the UEs to connect to both LTE and 5G networks simultaneously. This feature is specifically beneficial for 5G deployments at higher frequencies that experience significant signal blockage, since UEs can latch with 4G Macro Base station and use 5G layer capacity as and when it is available. In 3GPP standardization, 5G Multi-Connectivity is introduced in many configurations [4]. But one configuration, that is most likely to be introduced in the initial phases of 5G deployment is EN-DC (E-UTRAN New-Radio - Dual Connectivity). This is an extension of LTE Dual-Connectivity feature for a network with a mix of LTE and 5G-NR systems. To maintain multiple simultaneous connections with UE across different radios cells, the network must support stringent high bandwidth and low-latency links (mid-haul links) between the BTS nodes.

Towards addressing capacity requirements, the MNOs note the high variation of radio capacity requirements at different time of the day. Thus, they are considering deployment of RAN function in the Cloud infrastructure as a step towards energy efficient 5G deployment. As part of this setup, a part of BTS processing functions in the Baseband Units (BBU) is virtualized and moved to the Cloud component closer to the radio site (known as Edge Cloud). Thus, support of advanced LTE and 5G-NR features puts specific requirement on communication mechanisms between different Virtual Machines (VM) in the Cloud Infrastructure.

In this paper, we analyze the characteristics of the mid-haul link for LTE-NR Multi-Connectivity feature (specifically E-UTRAN-NR Dual Connectivity EN-DC configuration) in a heterogeneous C-RAN deployment. We simulate the C-RAN scenario using NS3 network simulator with heterogeneous mix of 4G and 5G-NR Base stations with appropriate cloud architecture split. We use the BTS power model to derive the Computational Requirement (CR) of the Base station for vBBU (virtual Baseband Unit) placement algorithm in C-RAN. Through simulation we review, for EN-DC configuration, the bandwidth requirement of the mid-haul link and assess variation in the end-to-end delay when mid-haul link latency is varied. We observe that, compared to traditional X2 link (whose bandwidth was a small fraction of S1 User plane traffic), the bandwidth of mid-haul link between LTE and 5GNR Base station scales with the data traffic of the EN-DC UEs in the Macro Base station sector. Further, we observe that the end-to-end delay scales with mid-haul link latency. From available cloud infrastructure bench-mark reports, we observe that the latency of inter VM communication links, varies depending on whether the two communicating VMs are on same or different Cloud servers. Thus, we propose "Neighbour Association-aware Placement" (NAP) algorithm for placement of the vBBUs in the same Cloud server and assess the benefits for EN-DC configuration.

II. FORMULATION OF SYSTEM

Fig. 1 shows the system model we have considered in this paper. It consists of a Mobile Network with Cloud based RAN (C-RAN). The C-RAN contains the Base Band Unit (BBU) processing (which we term as Cloud Base station) at a centralized location, the Remote Radio Unit (RRU) is at Radio site. The Cloud Base station is implemented with general-purpose processors to address the on-demand computing requirements. In the overall end to end Cloud deployment, C-RAN forms part of the Edge Cloud. At the Edge Cloud, Mobile Edge Orchestrator (MEO) takes care of configuration of the Cloud Infrastructure. This may be controlled by Software Defined Networking (SDN) Controller which has complete view of the all nodes of the mobile network. C-RAN architecture has several protocol layer-splits for placement of different Radio functions at different nodes in the network [3]. The protocol layer-split, may be decided based on (a) present RAN installed based with the MNO, and (b) the fronthaul capacity and latency offered by the transport network. The C-RAN model being considered in this paper is shown in Fig. 1. The C-RAN has two parts, (a) Cloud Base stations and (b) Mobile Edge Orchestrator (MEO). The Cloud Base station offers processing, and storage resources for Cloud Base station/Edge Computing. The MEO takes care of the configuration of Cloud Base stations. Cloud Base stations have the virtual BBUs (vBBU), which are virtualized form of the Base station BBU.

A. E-UTRAN NR-Dual Connectivity

E-UTRAN NR-Dual Connectivity (EN-DC) is a 3GPP standardized feature [4], that allows 4G and 5G coexistence

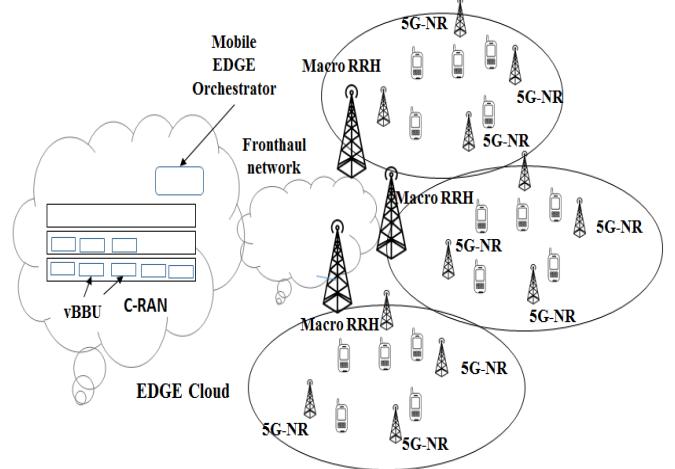


Fig. 1. Cloud based heterogeneous RAN with LTE and 5G-NR

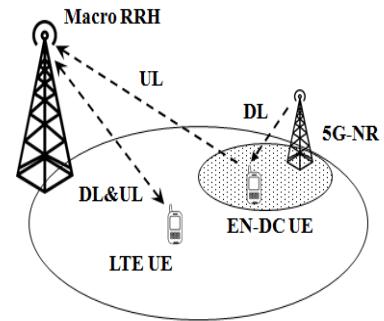


Fig. 2. E-UTRAN-NR Dual Connectivity (EN-DC)

and provides for smooth transition option for MNOs towards 5G. As shown in Fig. 2, the EN-DC UEs have connections with both LTE and 5G cells. Optionally, the EN-DC UE may also opt for supporting Downlink through 5G cell and Uplink through LTE. This enhances the uplink coverage for the UE since LTE operates at lower frequency band.

While a UE enters coverage of Macro BTS, it is attached to Macro LTE BTS, but it latches to 5G cells as well, if available. Connecting with the 5G-NR cell allows fast switching to avoid interruptions. Once the link is switched, the data-path is anchored in Macro BTS and the data traffic is routed to respective 5G-NR cell over Inter BTS mid-haul link. The switching of the cells happens dynamically and based on channel measurements. Thus availability of high capacity and low latency Inter-BTS (mid-haul) link is necessary for this feature to work.

B. Inter Base station mid-haul links

Traditionally, mid-haul links have been configured considering UE's connect-mode handover traffic [5]. The capacity requirement was thus derived as a fraction(< 5%) of the S1-U traffic (which is low). Further, traditional BTS's neighbour association (neighbour cell or adjacency list) has been decided

considering handovers only and has been mainly based coverage overlap of the neighbouring cells.

With features like EN-DC requiring higher bandwidth and low latency mid-haul links, the Inter BTS neighbour association needs to be re-assessed. The Inter BTS association can be represented as Graphs considering each BTS as a vertex and inter BTS (mid-haul) links as edges of the Graph). In Heterogeneous Network with support for data-path Multi-Connectivity feature, we would require only subset of existing adjacency matrices for inter BTS mid-haul links.

As per [6], which discusses different fronthaul transport requirements in case of Cloud RAN deployment with split protocol layers, the capacity requirement of the fronthaul link when the protocol function split is performed at MAC is given by (1).

$$R = N_L N_{SC,act} N_S R_{C,MCS} \log_2 M_{MCS} \mu \gamma T_{SF}^{-1} \quad (1)$$

where N_L corresponds to number of layers supported in the cell, $N_{SC,act}$ is the number of active subcarriers in the cell, N_S is the number of symbols in a subframe, $R_{C,MCS}$ is the coding rate used for transmission, $\log_2 M_{MCS}$ modulation order, μ is the load in the cell, γ is the overhead factor and T_{SF} is subframe duration.

C. NAP Algorithm for EN-DC in Cloud RAN

The placement of vBBUs into minimal number of Cloud servers, is envisaged to be a function of MEO. It is formulated as a bin-packing problem with a set of constraints. Since bin-packing problem is NP-Hard, it is normally solved through different heuristic methods. e.g. First-Fit Decreasing, Next-Fit, to maintain highest utilization at each of the Cloud servers that are active. Though these approaches work satisfactorily in normal cloud server operations, we need to review their benefits while supporting multi-cell coordination features like EN-DC that may require high level of communication between the VMs implementing vBBU.

From Cloud Infrastructure bench-marking report [7], we note that there is a high variation in the latency for VM to VM communication under different configurations (when communicating VMs are placed on the same Cloud server, and when they are placed on other cloud servers). We note that the delay is arising out of the volume of data traffic, and delays in intermediary switch and routers that are placed to inter-connect the servers in the cloud infrastructure. Hence, we observe that appropriate placement of VMs that have high capacity mid-haul requirements may be important.

In this paper, we propose and assess benefits of an NAP algorithm that considers the proximity of vBBUs that are involved in Multi-Connectivity configuration, while placement the vBBUs on Cloud server. We assume that

- 1) each vBBU maps to a single cell (either Macro, or 5G-NR) at the Radio site.
- 2) vBBU is an atomic processing unit and cannot be mapped to more than 1 Cloud server.

- 3) vBBUs requiring low latency and high capacity link between them (EN-DC configuration) are grouped and placed together.

Virtualization of BBU requires breaking down of the RAN functions and assigning equivalent computation complexity. We derive this in the next section.

D. Computational Resource Model for vBBU

In this section, we derive the equations for Computational Resource (CR) for the vBBU for both LTE and 5G-NR. We first compute the overall processing capacity of Cloud server which is as follows:

$$CN_{capacity} = N_{CPU} f N_{cores} N_{Inst. per cycle} \quad (2)$$

where N_{CPU} is CPU count per Cloud server, f is the CPU frequency in MHz, N_{cores} is the number of cores per CPU and $N_{Inst. per cycle}$ is CPU instructions per cycle. The CR that is needed by each vBBU, can be derived using the power model in [8] and [9]. The power model provides the reference values and scaling factor of different digital components of the BTS BBU. The Computational Resource (CR) of the digital components is determined in units of GOPS (Giga Operations Per Second). The CR derivation for vBBU is defined below.

For each cell, the processing requirement for BBU is given by

$$CR_{BBU \text{ per cell}} = (CR_{Static} + CR_{load-dependent}) \quad (3)$$

where CR_{Static} is independent of the load in the cell and the $CR_{load-dependent}$ scales with number of active mobile users or UEs in the cell.

In this paper, we consider lower-layer1 split for the front-haul (viz. split I_U and II_D) as given in [10]. Thus, we consider the CR_{Static} to be kept at the Radio site and the $CR_{load-dependent}$ to be moved to the cloud as vBBU. The BBU processing of vBBU can be expanded as following:

$$CR_{vBBU} = CR_{u,t} = \sum_{u \in U} \sum_{i \in I} CR_{u,i,ref} \prod_{x \in X} \left(\frac{x_{act}}{x_{ref}} \right)^{s_{i,x}} \quad (4)$$

where $i \in I$ are the sub-components, $x \in X$ are the input/scaling parameters [8], and $u \in U$ are the mobile users in the cell. From [8], we consider Macro Base stations as under Large Base station category, and 5G-NR Base station in LSAS or small cell category. Since we have a mix of Base station types in our setup viz. (a) Macro RRU, and (b) 5G-NR, we split the equations as below.

$$\begin{aligned} CR_{u,t}^{Macro_{dl}} = & CR_{MIMO \text{ Precoding}} + CR_{OFDM \text{ Mod-Demod}} \\ & + CR_{Mapping \text{ Demapping}} + CR_{Network} \\ & + CR_{Channel \text{ Coding}} \end{aligned} \quad (5)$$

$$\begin{aligned} CR_{u,t}^{5G-NR_{dl}} = & CR_{MIMO \text{ Precoding}} + CR_{OFDM \text{ Mod-Demod.}} \\ & + CR_{Mapping \text{ Demapping}} + CR_{Channel \text{ Coding}} \\ & + CR_{Network} \end{aligned} \quad (6)$$

$$\begin{aligned} CR_{u,t}^{Macro_{ul}} = & CR_{ChannelEst.} + CR_{OFDMMod-Demod.} \\ & + CR_{MappingDemapping} + CR_{Network} \\ & + CR_{ChannelCoding} + CR_{EqualizerComp.} \\ & + CR_{Equaliz.} \end{aligned} \quad (7)$$

$$\begin{aligned} CR_{u,t}^{5G-NR_{ul}} = & CR_{MIMO \text{ Precoding}} + CR_{ChannelCoding} \\ & + CR_{MappingDemapping} + CR_{Network} \\ & + CR_{OFDMMod-Demod.} \end{aligned} \quad (8)$$

The total Computational Resource requirement at the edge cloud (across all the cells) considering both uplink and down-link components is given by

$$CR_{CloudBase \text{ station}} = \sum_{j=1}^J \sum_{u \in U} CR_{u,t}^{Macro/5G-NR} \quad (9)$$

where $i \in I$ are the sub-components, and $j \in J$ are the cells in the cluster. We include a factor of 0.001 to (5) - (9), to address its computation at each scheduling interval (of 1ms). The CR_{vBBU} is collated over a period of 1 second for further analysis for the placement algorithm below.

For evaluating the benefits of the NAP algorithm, we keep note of the power consumption associated with Cloud Base station. $P_{Cloud \text{ base station}}$ is given by

$$P_{Cloud \text{ base station}} = \sum_{m=1}^M P_m^{CN} \quad (10)$$

where P_m^{CN} the power consumption at the Cloud server m . Assuming linear power model for the Cloud server, the power consumption at a given CPU utilization of the Cloud server is given by

$$P_m^{CN} = P_0^{CN} + \Delta_P^{CN} \rho_{CN} P_{max}^{CN} \quad (11)$$

where P_0^{CN} and P_{max}^{CN} denote the power consumption of the Cloud server when in idle mode and under full load respectively. Δ_P^{CN} denotes the slope of the equivalent linear power model and ρ_{CN} denotes the CPU utilization of the Cloud server.

III. SIMULATION AND RESULTS

A. Simulation Setup

For our simulation, we use mmWave [11] and LTE [12] modules of NS3 network simulator. The NS3-mmWave package, provides support for 5G Multi-Connectivity as one of the features. In the simulation setup we create Macro BTS with 5G-NR cells in a Heterogeneous Network layout. The

LTE Macro and 5G-NR BTS are connected through peer-to-peer link that simulates a X2 connectivity (mid-haul link) with a given bandwidth. We configure the simulation setup as per 3GPP Air Interface Specifications [13], and [14]. For 5G-NR, we consider 28GHz operation with maximum bandwidth of 400 MHz (as standardized by Release 15 of 3GPP Specifications). For the TDD frame format for 5G-NR, we configure the setup as per the 3GPP specifications. Hence, each 5G-NR frame is of 10ms and has 10 sub-frames. Each sub-frame is split into 8 slots each with 14 symbols.

TABLE II
RADIO NETWORK PARAMETERS

Parameters	Values
Bandwidth	LTE 20 MHz, 5G-NR 400 MHz
Number of UEs in cell	Up to 5 UEs and high throughput download (one UE with constant speed mobility)
UE Traffic model	Constant rate UDP traffic (Packet size 1024 with packet interval of 20 μs).
MIMO Mode in each Cell	Single user MIMO with single layer
UE Modulation Range	QPSK-64QAM
EN-DC Configuration (per sector)	1 eNB with 4 5G-NR units , UE Uplink on LTE
Number of EN-DC HetNets in a cluster	10 Radio sites each with 3 sectors in EN-DC Configuration
Packet Scheduler Algorithm	Proportionate Fair (PF)
BTS power	Macro, 5G-NR 40W per sector
Simulation time	10 seconds

TABLE III
CLOUD SERVER PARAMETERS

Parameters	Values
Number of cores per CPU N_{cores}	18
Number of CPUs per Cloud server	2
Cloud server Idle Power P_0^{CN}	180 W
Cloud server Maximum Power P_{max}^{CN}	543 W
Server power gradient ρ_P^{CN}	0.67
Base processor frequency f	2.3 GHz
Number of Instructions per CPU cycle $N_{Inst. \text{ per cycle}}$	16
Cloud server GOPS CN_{cap}	1324

With the NS3 mmWave package, we simulate radio network cluster of 10 sites each with 3 sectors. Each sector has 1 Macro LTE cell with 4 5G-NR BTS providing hot-spots. In each sector, we simulate up to 5 UE with heavy download. Out of the 5 UEs in the sector, one UE has with mobility with constant speed. We consider constant bitrate UDP traffic for each of these UEs in the cell in both uplink and downlink directions. The NS3 radio network configuration parameters are listed in Table II.

Within the NS3 network simulator, we define a Mobile Edge Orchestrator (MEO) unit. This MEO regularly monitors the CR_{vBBU} of each vBBU registered with the Cloud Base station. Based on latest estimates on the CR_{vBBU} , MEO computes the best placement of the vBBU in the given set

of active Cloud servers. We implement the NAP algorithm in this unit and simulate migration of VMs across Cloud servers by appropriately setting the delay of the inter BTS mid-haul link to that of inter Cloud server delay. The Cloud server parameters are derived from [15] and [16] and are listed in Table III. In the simulation, we monitor the following:

- Radio resource allocations for each UE at each scheduling interval.
- Bandwidth and latency on inter-BTS X2 link in case of EN-DC configuration.
- Delay experienced by packets (measured between LTE eNodeB PDCP layer and UE PDCP layer).
- Cloud Base station server occupancy and power consumption.

B. Results and Inferences

1) *X2 link capacity for EN-DC*: Here, we analyze the capacity requirement for mid-haul link between LTE and NR BTS for Multi-Connectivity.

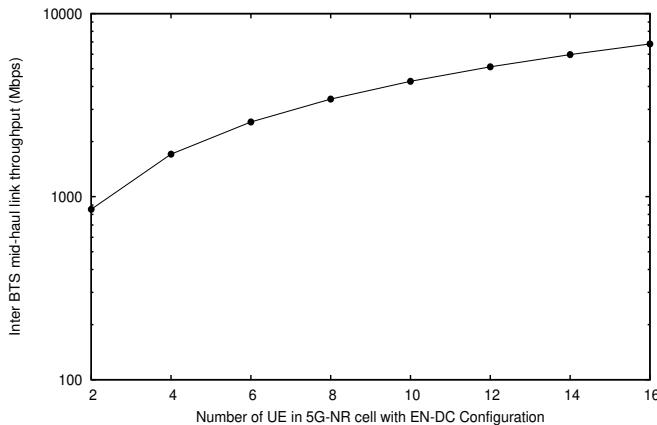


Fig. 3. Inter BTS mid-haul link throughput variation

Fig. 3 highlights the variation of mid-haul link capacity when number of EN-DC UEs are increased in the 5G-NR cell. We see that X2 link capacity increases proportionally with increase in load in the 5G-NR cell. This is also reflected from (1). Since the 5G-NR cell supports bandwidth upto 400MHz (and higher bandwidth in future), the results indicates the range of bandwidth that may be needed to support this feature. We further note that the Inter BTS mid-haul link capacity requirement was $\sim 4\text{--}5\%$ more than the throughput supported by it at the application (owing to the overheads).

2) *Variation of PDCP delay with X2 link latency*: Here, we analyze the variation in the end to end delay observed at PDCP layer when the X2 link latency is varied. For this, we considered a single EN-DC UE case. The UE is connected to both Macro and 5G-NR cells and is moving with constant speed. Thus, it experiences 5G-NR cell switch-over when the signal quality degrades. The switch-over decision is based on measurements shared over X2. Fig. 4 highlights the variation of delay observed at PDCP layer as a function of X2 link latency (or inter VM delay in the Cloud Base station). We

observe that the delay measured at PDCP increases proportionally with the VM to VM (or X2 link) delay. We further note that with higher VM to VM delay, the end to end delay at PDCP is quite observable. The higher VM to VM delay also impacted the throughput due to delayed switch-over decisions.

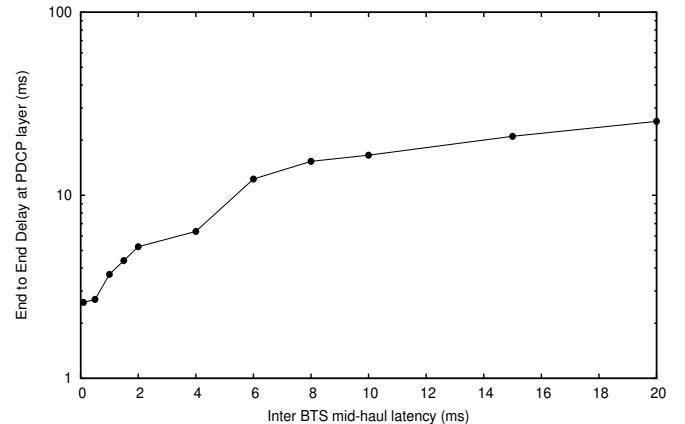


Fig. 4. Delay at PDCP with variation in latency at Inter-BTS mid-haul link

3) *Effect of neighbour association-aware vBBU placement*: Here, we analyze the gain due to the Neighbour association-aware vBBU placement algorithm. In this simulation, we consider a cluster of 10 sites, each with 3 sector with Macro BTS with 4 5G-NR units, with 5 UEs in each sector. We then use NAP algorithm (where the HetNet BTSs of given sector are always placed together in the cloud server) for vBBU mapping to cloud server and compared it with the traditional "First-Fit-Decreasing" algorithm. We repeated the simulation over 10 iterations, with varying position of UEs and latching them to the nearest BTS each time.

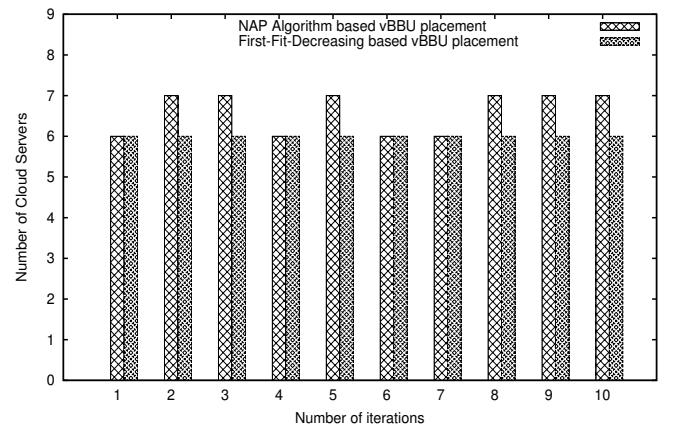


Fig. 5. Comparison of Cloud server count for NAP and First-fit-Decreasing algorithms

Fig. 5 highlights the number of Cloud server that are used for given C-RAN cluster. We note that, due to additional constraint of grouping the HetNet BTS (Macro eNodeB and its associated 5G-NR cells) together, NAP algorithm consumed

extra Cloud Server on many occasions, whereas “First-Fit-Decreasing” algorithm, maintained a lower count of number of Cloud Servers. But since we assume linear power model, the difference in power consumption was not significant.

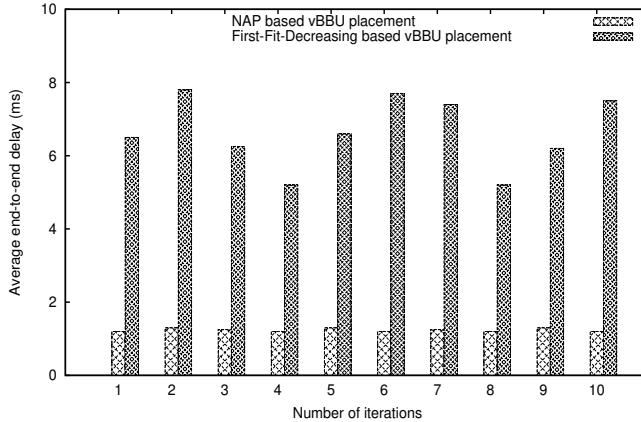


Fig. 6. Comparison of end-to-end latency for NAP and First-fit-Decreasing algorithms

We further note that, in case of NAP algorithm, the end-to-end latency at PDCP layer was consistent and lower compared to “First-Fit-Decreasing” algorithm, as shown in Fig. 6. In the EN-DC simulation setup, the uplink and downlink traffic was handled by the Macro and 5G-NR Base stations respectively. The CR associated with 5G-NR Base stations were higher than the Macro Base stations. Thus, the “First-Fit-Decreasing” algorithm placed all the 5G-NR vBBUs first before placing Macro vBBU. This resulted in Macro vBBUs and its corresponding 5G-NR vBBUs (part of same sector) being placed on different cloud server in all the iterations, thus resulting in higher end-to-end delay. In case of NAP algorithm, the vBBUs participating in the Multi-Connectivity or EN-DC were always placed together. Hence it resulted to minimal end-to-end delay on the downlink. Thus, NAP algorithm proved to be beneficial in EN-DC configuration specifically for delay sensitive applications by offering higher capacity and maintaining lower end-to-end delay.

IV. CONCLUSION

In this paper, we analyzed the characteristics of the mid-haul links for LTE-NR Multi-Connectivity feature (specifically for E-UTRAN-NR Dual Connectivity configuration) in a heterogeneous C-RAN deployment. We simulated the C-RAN scenario with NS3 network simulator with heterogeneous mix of 4G and 5G-NR Base stations with appropriate protocol architecture split. We used the BTS power model to derive the Computational Requirement (CR) of the Base station sub-components for vBBU placement algorithm. Through simulations, we observed that for the EN-DC case, throughput requirement of inter BTS mid-haul link is proportional to forwarded user traffic that it carries, with $\sim 4 - 5\%$ overhead. The overall bandwidth depends on the maximum bandwidth that can be supported by the UE and the target cell. We further

observed that end-to-end delay is sensitive to inter BTS mid-haul link. We, thus, proposed ”Neighbour Association-aware Placement” (NAP) algorithm for placement of the vBBUs in the Cloud server. This algorithm keeps track of the vBBUs that interact with each other for EN-DC feature and hence, places them together on the same Cloud Server to avoid the inter VM latency and thus maintains lower delay at PDCP. We find that the NAP algorithm proved to be beneficial in EN-DC configuration specifically for delay sensitive applications by offering higher capacity and maintaining lower end-to-end delay.

REFERENCES

- [1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20172022, <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.pdf>, accessed 23 February 2018
- [2] Nokia Mobile Broadband India Traffic Index (MBIT index) 2018, <https://networks.nokia.com/in/mbit-index>, accessed 23 February 2018
- [3] Study on new radio access technology: Radio access architecture and interfaces (Release 14), 3GPP TR 38.801, v14.0.0, March 2017.
- [4] E-UTRA and NR; Multi-connectivity; Overall description; Stage-2 (Release 15), 3GPP TS 37.340, v15.2.0, June 2018.
- [5] Backhaul Provisioning for LTE-Advanced & Small Cells, https://www.ngmn.org/fileadmin/user_upload/150929_NGMN_P-SmallCells_Backhaul_for_LTE-Advanced_and_Small_Cells.pdf, accessed 01 October 2018.
- [6] Bartelt J, Vucic N, Camps-Mur D, et al, “5G transport network requirements for the next generation fronthaul interface,” *EURASIP Journal on Wireless Communications and Networking*, 2017 (1) 89.
- [7] Intel Open Network Platform Release 2.1 Performance Test Report, https://download.01.org/packet-processing/ONPS2.1/Intel_ONP_Release_2.1_Performance_Test_Report_Rev1.0.pdf, accessed 10 September 2018
- [8] B. Debaillie, C. Dessel, and F. Louagie, “A Flexible and Future-Proof Power Model for Cellular Base Stations,” *IEEE 81st Vehicular Technology Conference*, 2015, pp. 17.
- [9] C. Dessel, B. Debaillie and F. Louagie, “Modeling the hardware power consumption of large scale antenna systems,” *IEEE Online Conference on Green Communications (OnlineGreenComm)*, Tucson,AZ, 2014, pp. 1-6.
- [10] eCPRI - Common Public Radio Interface Specification, V1.0, August 2017.
- [11] M. Mezzavilla, M. Zhang,M. Polese et.al,“End-to-End Simulation of 5G mmWave Networks,” *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, 2018
- [12] G. Piro, N. Baldo, and M. Miozzo., “An LTE module for the ns-3 network simulator,”In Proceedings of the 4th International ICST Conference on Simulation Tools and Techniques, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, Belgium, 2011, pp. 415-422.
- [13] NR - Base Station (BS) Radio Transmission and Reception, 3GPP TS 38.104, v15.2.0, June 2018.
- [14] NR - Physical channels and modulation, 3GPP TS 38.211, v15.2.0, June 2018.
- [15] Intel Xeon Processor E5-2697 v4, https://ark.intel.com/products/91755/Intel-Xeon-Processor-E5-2697-v4-45M-Cache-2_30-GHz, accessed 28 October 2017.
- [16] Intel Xeon E5-2600 v4 Broadwell-EP Review, <http://www.tomshardware.co.uk/intel-xeon-e5-2600-v4-broadwell-ep/review-33513-8.html>, accessed 10 September 2018.

ASER Analysis of General Order Rectangular QAM for Dual-Hop NLOS UV Communication System

Kamal K. Garg, Praveen Kumar Singya, Vimal Bhatia

Discipline of Electrical Engineering, IIT Indore, India 453552

{phd1701102008, phd1501102023, v.bhatia}@iiti.ac.in

Abstract—Ultraviolet (UV) communication is capable of providing non-line-of-sight (NLOS) wireless connectivity due to strong molecular and aerosol scattering experienced at UV wavelength. The effect of atmospheric turbulence in NLOS UV channel is usually ignored under the assumption of short distance communication. However, for long distance, relay assisted communication is commonly used. In this work, we consider a dual-hop amplify-and-forward (AF) relayed outdoor NLOS UV communication system experiencing atmospheric turbulence, and derive closed-form expression of the average symbol error rate (ASER) for general order rectangular quadrature amplitude modulation (RQAM) scheme. The numerical values of ASER expression are compared with computer simulations to validate the accuracy of the theoretical analysis.

Index Terms—NLOS UV, amplify-and-forward (AF), ASER, RQAM

I. INTRODUCTION

Ultraviolet (UV) communication is an emerging optical wireless communication technology which has gained attention due to the advancements in solid state technologies. UV communication offers many advantages like ability to provide wide coverage, security and seamless connectivity as compared to radio-frequency (RF) communication [1]. In addition, UV communication can be used in scenarios where RF communication is prohibited, for example, in aircrafts and hospitals. UV communication system operates in solar blind deep UV band (wavelength 200-280 nm) which experiences extremely low background noise due to the absorption of UV solar blind radiation by ozone layer. In deep UV band, the atmospheric scattering becomes extremely high resulting in UV communication systems to operate in non-line-of-sight (NLOS) mode. Due to its NLOS nature, UV channel does not need line-of-sight (LOS) connectivity, thus, relaxing the pointing and tracking requirements of conventional optical communication link. However, NLOS connectivity results in very high attenuation in UV channels as compared to LOS links. Further, as the communication range increases, UV channel starts experiencing fading due to fluctuations in the refractive index of the atmosphere, thereby, deteriorating the performance of UV communication system [2].

A powerful solution to combat high path-loss and fading in UV communication is to use relay. Use of relays has been shown to enhance the spectral efficiency, coverage and reliability of the UV communication link [3, 4]. Moreover, a NLOS UV communication system with relays can be used for long distance communication. There are several relaying techniques proposed in the literature, out of which, the commonly used schemes are decode-and-forward (DF) and amplify-and-forward (AF). AF relaying has gained significant attention due to its computational simplicity, easy deployment and low energy requirements [5, 6]. Recently, there are several research studies reported in the literature, examining the performance of relay assisted UV communication systems [3, 4, 7, 8]. In [3], the outage analysis of DF relayed UV communication system is conducted for turbulent channel. In [4] error rate performance of OFDM based UV communication system using AF and DF relaying is studied; however, the effect of turbulence is ignored under the assumption of short distance communication. In [7], the authors proposed a cooperative reception technique for serially relayed NLOS UV link, neglecting the effect of turbulence.

In optical wireless communication (OWC) systems, on-off-keying (OOK) is the commonly used modulation technique, due to its simplicity [2, 9]. One of the downside of OOK modulation is its poor spectral efficiency. On the other hand, quadrature amplitude modulation (QAM) is a promising modulation technique, widely used in wireless communication systems due to its high spectral efficiency. Rectangular QAM (RQAM) is a variant of QAM which has got significant attention due to its generic nature [6, 10]. RQAM is a versatile modulation scheme which includes binary phase shift keying (BPSK), quadrature phase-shift keying (QPSK), square QAM (SQAM), non-square QAM and orthogonal binary frequency-shift keying (OBFSK) [6].

In this paper, we derive the closed-form expression of the average symbol error rate (ASER) for general order RQAM in a dual-hop AF relayed NLOS UV communication system experiencing weak atmospheric turbulence. Both the path losses due to scattering [11] and scintillation [12] are taken into account for the analysis. We present the lower-bound expression of the

outage probability, and compute the probability density function (PDF) of the end-to-end signal-to-noise (SNR) ratio of the link. The computed PDF is used to derive the average symbol error rate (ASER) expression of the system for general order RQAM. Performance of different RQAM constellations is studied and useful insights are drawn. It is shown that SQAM performs better than non-square RQAM for all constellation sizes. Correctness of the derived ASER expression is validated with the help of simulations. To the best of authors' knowledge, the analytical expressions derived in this work are novel and not available in the literature.

Rest of this paper is organized as follows: Section II and III describe the NLOS UV channel model and dual-hop AF relayed NLOS UV system model, respectively. In Section IV, lower-bound on the outage probability is presented. The ASER of the system for RQAM is derived in Section V. Section VI shows the numerical results to validate the accuracy of the derived analytical expressions. The paper finally concludes in Section VII. *Notations:* The following notations are used throughout the paper: $\ln(\cdot)$ represents natural logarithm. $\mathcal{N}(\mu, \sigma^2)$ denotes Gaussian random variable (RV) with mean μ and variance σ^2 . $\text{LogN}(\mu, \sigma)$ is the distribution of a lognormal RV, whose natural logarithm is $\mathcal{N}(\mu, \sigma^2)$. $\mathcal{Q}(x) = \frac{1}{2\pi} \int_x^\infty e^{-t^2/2} dt$. $\Pr(\cdot)$ denotes probability. $f_X(\cdot)$ and $F_X(\cdot)$ represent the probability density function (PDF) and cumulative distribution function (CDF) of a RV X , respectively. $\mathbb{E}[\cdot]$ is statistical expectation operator. Superscripts $(\cdot)^{Tx}$ and $(\cdot)^{Rx}$ denote transmitter and receiver, respectively.

II. CHANNEL MODEL

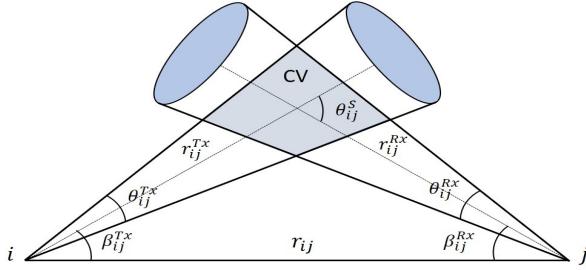


Fig. 1: NLOS UV Communication Channel

Fig. 1 shows a NLOS UV communication channel between node i and node j which are r_{ij} distance apart. In a dual-hop relayed communication system $i, j \in \{S, R, D\}$, $i \neq j$, where S , R and D represent source, relay and destination nodes, respectively. Node i acts as transmitter with elevation angle β_{ij}^{Tx} and beam divergence angle θ_{ij}^{Tx} . Node j acts as receiver with elevation angle β_{ij}^{Rx} and field of view (FOV) θ_{ij}^{Rx} . Let c_{ij} represents the channel coefficient of the communication link, and is modelled as

$$c_{ij} = \ell_{ij} h_{ij}, \quad (1)$$

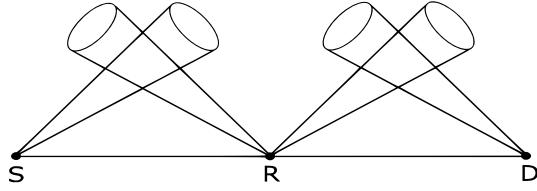


Fig. 2: Dual-hop NLOS UV communication system

where ℓ_{ij} is the path-loss component. ℓ_{ij} is the product of the attenuations due to scattering and turbulence, which are represented by ℓ_{ij}^S and ℓ_{ij}^T , respectively. ℓ_{ij}^T is computed as [3]

$$\ell_{ij}^T = 10^{-a/10}, \quad (2)$$

$$a = 2\sqrt{23.17C_n^2(2\pi/\lambda)^{7/6}} \left(\sqrt{(r_{ij}^{Tx})^{11/6}} + \sqrt{(r_{ij}^{Rx})^{11/6}} \right),$$

with λ representing the wavelength and C_n^2 is the structure parameter of refractive index fluctuation. r_{ij}^{Tx} and r_{ij}^{Rx} are the distances from node i to common volume (CV) and CV to node j , respectively, and are calculated as $r_{ij}^{Tx} = r_{ij} \sin(\beta_{ij}^{Rx}) / \sin(\beta_{ij}^S)$, $r_{ij}^{Rx} = r_{ij} \sin(\beta_{ij}^{Tx}) / \sin(\beta_{ij}^S)$, where $\sin(\beta_{ij}^S)$ is the scattering angle, $\sin(\beta_{ij}^S) = \sin(\beta_{ij}^{Tx}) + \sin(\beta_{ij}^{Rx})$. Expression for the path-loss due to scattering, ℓ_{ij}^S , is given in (6) on the next page. Parameters A_r , k_s , k_e and $\Psi(\cdot)$ are area of receiving aperture, scattering coefficient, extinction coefficient and scattering phase function, respectively; these parameters are defined in [11].

The channel is assumed to experience atmospheric turbulence which is the result of refractive index variations due to the change in atmospheric pressure and temperature. Parameter h_{ij} in (1) represents the fading channel coefficient due to turbulence and is modelled using lognormal distribution for weak/moderate turbulence, and using gamma-gamma distribution for strong turbulence [13, 14]. The severity of the turbulence induced fading is measured in terms of irradiance variance which is defined as $\sigma_{LOS}^2 = 1.23C_n^2 \left(\frac{2\pi}{\lambda} \right)^{7/6} r^{11/6}$ for plane wave propagation and LOS scenario [15]; here r is the distance between transmitter and receiver. In this study, we have considered the $\lambda = 260$ nm and $C_n^2 = 5 \times 10^{-15} \text{ m}^{-2/3}$ for which the weak turbulence condition is met for distance up to ~ 1 Km.

For weak turbulence, the PDF of the fading channel coefficient, h_{ij} , is derived in [3] as

$$f_{h_{ij}}(h) = \frac{1}{\sqrt{2\pi}\sigma_{ij}h} \exp \left(\frac{-(\ln(h) + \mu_{ij})^2}{2\sigma_{ij}^2} \right), \quad (3)$$

where μ_{ij} and σ_{ij}^2 are the mean and variance of $\ln(h_{ij})$, and are given as

$$\begin{aligned} \sigma_{ij}^2 &= 1.23C_n^2(2\pi/\lambda)^{7/6} r_{ij}^{11/6} \\ &\times \left(\frac{\sin(\beta_{ij}^{Tx})^{11/6} + \sin(\beta_{ij}^{Rx})^{11/6}}{\sin(\beta_{ij}^S)^{11/6}} \right), \end{aligned} \quad (4)$$

$$\mu_{ij} = 0.5\sigma_{ij}^2. \quad (5)$$

$$\ell_{ij}^S = \frac{96r_{ij} \sin(\beta_{ij}^{Tx}) \sin^2(\beta_{ij}^{Rx})(1 - \cos(\theta_{ij}^{Tx}/2)) \exp(k_e r_{ij} [\sin(\beta_{ij}^{Tx}) + \sin(\beta_{ij}^{Rx})]/\sin(\beta_{ij}^S))}{k_s A_r \theta_{ij}^{Tx^2} \theta_{ij}^{Rx} \Psi(\beta_{ij}^S) \sin(\beta_{ij}^S) [12 \sin^2(\beta_{ij}^{Rx}) + \theta_{ij}^{Rx^2} \sin^2(\beta_{ij}^{Tx})]} \quad (6)$$

III. SYSTEM MODEL

We consider a dual-hop NLOS UV communication system using AF relaying, as shown in Fig. 2. In AF relayed systems, the data transmission involves two time phases. In the first time phase, S sends the signal to R and, in the second time phase, R amplifies the received signal and forwards the same to D . Let P_T be the total power budget and, P_S and P_R represent the power allocated to S and R , respectively. The light emitting diode (LED) at the transmitter is derived in the forward bias region by applying appropriate DC-bias to the real and imaginary parts of the modulated symbols. An appropriate DC-bias, moves the RQAM constellation points to the first quadrant.

During the first time phase, the data received at R is given as

$$y_{SR} = R_r \ell_{SR} h_{SR} \sqrt{P_s} s + n_R, \quad (7)$$

where R_r is the receiver responsivity and s represents the independent and identically distributed source symbol with average energy normalized to one, $\mathbb{E}[s] = 1$. n_R is the noise at R , and is modelled using additive white Gaussian noise (AWGN) with mean μ_R and variance σ_R^2 as in [3, 13, 16].

In the second time phase, R amplifies the received signal, y_{SR} , and forwards the same to D . The signal received at D is given as

$$y_{RD} = \frac{R_r \ell_{RD} h_{RD} \sqrt{P_r}}{\sqrt{\mathbb{E}[y_{SR}^2]}} y_{SR} + n_D, \quad (8)$$

where, $n_D \sim \mathcal{N}(\mu_D, \sigma_D^2)$ is the AWGN at D . The received power at R is normalized using $\mathbb{E}[y_{SR}^2]$. Without loss of generality s , h_{SR} , h_{RD} , n_R and n_D are assumed to be independent. Thus, we can write $\mathbb{E}[y_{SR}^2] = R_r^2 \ell_{SR}^2 h_{SR}^2 P_S + \sigma_R^2$. On substituting y_{SR} from (7) and, $\mathbb{E}[y_{SR}^2]$ into (8), and after mathematical simplifications, we get

$$y_{RD} = R_r^2 \sqrt{P_s} G \ell_{SR} \ell_{RD} h_{SR} h_{RD} s + R_r G \ell_{RD} h_{RD} n_R + n_D, \quad (9)$$

where $G = \sqrt{P_r / (R_r^2 \ell_{SR}^2 h_{SR}^2 P_s + \sigma_R^2)}$.

IV. OUTAGE PROBABILITY

This Section presents the lower-bound on the outage probability of dual-hop NLOS UV communication link. The instantaneous SNR of the $S \rightarrow R \rightarrow D$ link is computed using (9) as

$$\Gamma_{SRD} = \frac{\Gamma_{SR} \Gamma_{RD}}{\Gamma_{SR} + \Gamma_{RD}}, \quad (10)$$

with Γ_{SR} and Γ_{RD} representing the instantaneous SNRs of $S \rightarrow R$ and $R \rightarrow D$ links, respectively, and are defined as

$$\Gamma_{SR} = (R_r \ell_{SR})^2 \gamma_{SR} h_{SR}^2, \quad \Gamma_{RD} = (R_r \ell_{RD})^2 \gamma_{RD} h_{RD}^2, \quad (11)$$

where $\gamma_{SR} = P_S / \sigma_{SR}^2$ and $\gamma_{RD} = P_R / \sigma_{RD}^2$ are the transmit SNRs of $S \rightarrow R$ and $R \rightarrow D$ links, respectively. h_{SR} and h_{RD} are lognormally distributed fading coefficients of $S \rightarrow R$ and $R \rightarrow D$ links, respectively, and are defined as, $h_{SR} \sim \text{LogN}(\mu_{SR}, \sigma_{SR})$ and $h_{RD} \sim \text{LogN}(\mu_{RD}, \sigma_{RD})$. Using properties of lognormal distribution, parameters $\mu_{SR}, \sigma_{SR}, \mu_{RD}$ and σ_{RD} can be computed from (11) as

$$\begin{aligned} \mu_{\Gamma_{SR}} &= 2\mu_{SR} + \ln(R_r \ell_{SR})^2 \gamma_{SR}, & \sigma_{\Gamma_{SR}} &= 2\sigma_{SR}, \\ \mu_{\Gamma_{RD}} &= 2\mu_{RD} + \ln(R_r \ell_{RD})^2 \gamma_{RD}, & \sigma_{\Gamma_{RD}} &= 2\sigma_{RD}. \end{aligned} \quad (12)$$

The outage probability of $S \rightarrow R \rightarrow D$ link is defined as

$$\Pr(\Gamma_{SRD} < \gamma_{th}) = F_{\Gamma_{SRD}}(\gamma_{th}), \quad (13)$$

where, γ_{th} is a predefined threshold. From [17], the lower-bound on the outage probability is given as

$$\begin{aligned} P_{out}^{lb} &= \Pr(\Gamma_{SRD}^{ub} < \gamma_{th}) = F_{\Gamma_{SRD}^{ub}}(\gamma_{th}) \\ &= 1 - (1 - \mathcal{Q}(\Xi_{SR})) (1 - \mathcal{Q}(\Xi_{RD})), \end{aligned} \quad (14)$$

with $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-u^2/2} du$, Γ_{SRD}^{ub} denoting the upper-bound on Γ_{SRD} , $\Xi_{SR} = (\ln \gamma_{th} - \mu_{\Gamma_{SR}}) / \sigma_{\Gamma_{SR}}$ and $\Xi_{RD} = (\ln \gamma_{th} - \mu_{\Gamma_{RD}}) / \sigma_{\Gamma_{RD}}$. In medium to high SNR regime, the lower-bound on the outage probability can be approximated by

$$\tilde{P}_{out}^{lb} = \tilde{F}_{\Gamma_{SRD}^{ub}}(\gamma_{th}) = \mathcal{Q}(\Xi_{SR}) + \mathcal{Q}(\Xi_{RD}). \quad (15)$$

The corresponding expression for the PDF of $S \rightarrow R \rightarrow D$ link can be evaluated from (15) as

$$\begin{aligned} \tilde{f}_{\Gamma_{SRD}}(\gamma) &= \frac{d\tilde{F}_{\Gamma_{SRD}^{ub}}(\gamma)}{d\gamma} \\ &= \frac{1}{\sqrt{2\pi} \gamma \sigma_{\Gamma_{SR}}} \exp \left[-\frac{1}{2} \left(\frac{\ln \gamma - \mu_{\Gamma_{SR}}}{\sigma_{\Gamma_{SR}}} \right)^2 \right] \\ &\quad + \frac{1}{\sqrt{2\pi} \gamma \sigma_{\Gamma_{RD}}} \exp \left[-\frac{1}{2} \left(\frac{\ln \gamma - \mu_{\Gamma_{RD}}}{\sigma_{\Gamma_{RD}}} \right)^2 \right]. \end{aligned} \quad (16)$$

V. ASER ANALYSIS

The conditional SER expression of $M_I \times M_Q$ RQAM modulation scheme over AWGN channel is given by [10, 13, eq. (6)]

$$P_s(e|\gamma) = 2p\mathcal{Q}(a\sqrt{\gamma}) + 2q\mathcal{Q}(b\sqrt{\gamma}) - 4pq\mathcal{Q}(a\sqrt{\gamma})\mathcal{Q}(b\sqrt{\gamma}), \quad (17)$$

where $p = 1 - 1/M_I$, $q = 1 - 1/M_Q$, $a = \sqrt{6/((M_I^2 - 1)(M_Q^2 - 1)\beta^2)}$, $b = \beta a$ and $\beta = d_Q/d_I$, in which d_I and d_Q denote in-phase and quadrature decision distances, respectively.

Using (16) and (17), ASER can be evaluated as

$$\overline{P}_s = \int_0^\infty P_s(e|\gamma) \tilde{f}_{\Gamma_{SRD}}(\gamma) d\gamma \quad (18)$$

$$= \int_0^\infty [2p\mathcal{Q}(a\sqrt{\gamma}) + 2q\mathcal{Q}(b\sqrt{\gamma}) - 4pq\mathcal{Q}(a\sqrt{\gamma})\mathcal{Q}(b\sqrt{\gamma})] \\ \times \left[\frac{1}{\sqrt{2\pi}\gamma\sigma_{\Gamma_{SR}}} \exp\left[-\frac{1}{2}\left(\frac{\ln\gamma - \mu_{\Gamma_{SR}}}{\sigma_{\Gamma_{SR}}}\right)^2\right] \right. \\ \left. + \frac{1}{\sqrt{2\pi}\gamma\sigma_{\Gamma_{RD}}} \exp\left[-\frac{1}{2}\left(\frac{\ln\gamma - \mu_{\Gamma_{RD}}}{\sigma_{\Gamma_{RD}}}\right)^2\right] \right] d\gamma. \quad (19)$$

Let

$$I_1(\alpha_1, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty \frac{1}{\gamma} \mathcal{Q}(\alpha_1\sqrt{\gamma}) \\ \times \exp\left[-\frac{1}{2}\left(\frac{\ln\gamma - \mu}{\sigma}\right)^2\right] d\gamma \quad (20)$$

and

$$I_2(\alpha_1, \alpha_2, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty \frac{1}{\gamma} \mathcal{Q}(\alpha_1\sqrt{\gamma})\mathcal{Q}(\alpha_2\sqrt{\gamma}) \\ \times \exp\left[-\frac{1}{2}\left(\frac{\ln\gamma - \mu}{\sigma}\right)^2\right] d\gamma. \quad (21)$$

Using (20) and (21), (19) can be written as

$$\begin{aligned} \overline{P}_s = & 2p[I_1(a, \mu_{\Gamma_{SR}}, \sigma_{\Gamma_{SR}}) + I_1(a, \mu_{\Gamma_{RD}}, \sigma_{\Gamma_{RD}})] \\ & + 2q[I_1(b, \mu_{\Gamma_{SR}}, \sigma_{\Gamma_{SR}}) + I_1(b, \mu_{\Gamma_{RD}}, \sigma_{\Gamma_{RD}})] \\ & - 4pq[I_2(a, b, \mu_{\Gamma_{SR}}, \sigma_{\Gamma_{SR}}) + I_2(a, b, \mu_{\Gamma_{RD}}, \sigma_{\Gamma_{RD}})]. \end{aligned} \quad (22)$$

Integrals $I_1(\alpha_1, \mu, \sigma)$ and $I_2(\alpha_1, \alpha_2, \mu, \sigma)$ are evaluated in the Appendix. Substituting (26) and (27) into (22), and after simplifications, the final expression of ASER is computed and is given in (23).

VI. NUMERICAL AND SIMULATION RESULTS

In this Section, Monte Carlo simulations are performed to verify the analytical expression for ASER derived in Section V. We assume wavelength $\lambda = 260$ nm, scattering coefficient $k_s = 5.5 \times 10^{-4} \text{ m}^{-1}$, extinction coefficient $k_e = 1.352 \times 10^{-3} \text{ m}^{-1}$, receiver aperture area $A_R = 1.77 \text{ cm}^2$ and refractive index structure coefficient $C_n^2 = 5 \times 10^{-15} \text{ m}^{-2/3}$. Further, we assume source transmitter beam divergence $\theta_{SR}^{Tx} = 8 \text{ mrd}^1$, source elevation angle $\beta_{SR}^{Tx} = 45^\circ$, destination FOV $\theta_{RD}^{Rx} = 30$, destination elevation angle $\beta_{RD}^{Tx} = 45^\circ$, transmitter and receiver elevation angle at relay $\beta_{SR}^{Rx} = \beta_{RD}^{Tx} = 70^\circ$, and relay transmitter beam divergence $\theta_{RD}^{Tx} = 8 \text{ mrd}$. Furthermore, the distance between S and D , r_{SD} , is considered as 1 Km [3], and R is assumed to be placed in the middle of S and D . Both S and R are assumed to have equal transmit power $P_S = P_R = 0.5$, such that the total power budget, P_T , remains unity.

Fig. 3(a) - 3(c) depicts the ASER versus SNR performance curves for $S \rightarrow D$ and $S \rightarrow R \rightarrow D$ links, for different SQAM and RQAM constellations. It is observed that simulation results overlap with theoretical results confirming the accuracy of the derived analytical ASER expression. It is evident from the plots that the relayed NLOS link ($S \rightarrow R \rightarrow D$) outperforms the direct NLOS link ($S \rightarrow D$) for all the considered constellations.

Fig. 3(a) shows the ASER performance curves of RQAM for 32 points constellations. It is observed that 8×4 -RQAM performs better than 16×2 -RQAM. To

¹mrd stands for milliradian. 1 mrd = 0.001 radian.

$$\begin{aligned} \overline{P}_s = & \frac{1}{\sqrt{\pi}} \sum_{i=1}^n w_i \left[2p \left\{ \mathcal{Q}\left(\sqrt{a \exp(\sqrt{2}\sigma_{\Gamma_{SR}}x_i + \mu_{\Gamma_{SR}})}\right) + \mathcal{Q}\left(\sqrt{b \exp(\sqrt{2}\sigma_{\Gamma_{RD}}x_i + \mu_{\Gamma_{RD}})}\right) \right\} \right. \\ & + 2q \left\{ \mathcal{Q}\left(\sqrt{b \exp(\sqrt{2}\sigma_{\Gamma_{SR}}x_i + \mu_{\Gamma_{SR}})}\right) + \mathcal{Q}\left(\sqrt{a \exp(\sqrt{2}\sigma_{\Gamma_{RD}}x_i + \mu_{\Gamma_{RD}})}\right) \right\} \\ & - 4pq \left\{ \mathcal{Q}\left(\sqrt{a \exp(\sqrt{2}\sigma_{\Gamma_{SR}}x_i + \mu_{\Gamma_{SR}})}\right) \mathcal{Q}\left(\sqrt{b \exp(\sqrt{2}\sigma_{\Gamma_{SR}}x_i + \mu_{\Gamma_{SR}})}\right) \right. \\ & \left. \left. + \mathcal{Q}\left(\sqrt{b \exp(\sqrt{2}\sigma_{\Gamma_{RD}}x_i + \mu_{\Gamma_{RD}})}\right) \mathcal{Q}\left(\sqrt{a \exp(\sqrt{2}\sigma_{\Gamma_{RD}}x_i + \mu_{\Gamma_{RD}})}\right) \right\} \right] \end{aligned} \quad (23)$$

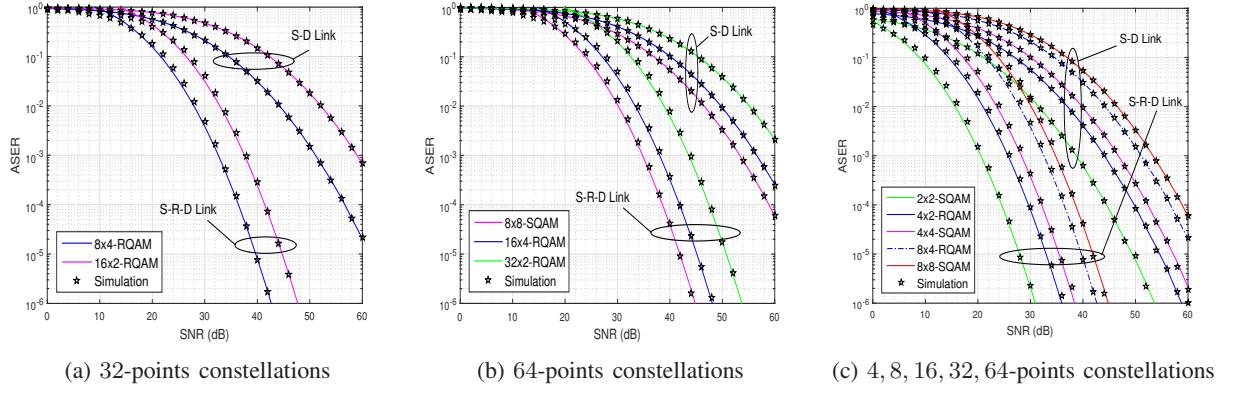


Fig. 3: Comparison of analytical and simulation results of ASER versus SNR for different RQAM schemes.

achieve the ASER of 10^{-4} , an SNR gain of approx. 6 dB is observed for 8×4 -RQAM as compared to 16×2 -RQAM. Further, referring to the ASER curves for 64-points constellations in Fig. 3(b), it is observed that 8×8 -SQAM provides a significant gain of 3 dB (approx.) and 8 dB (approx.) when compared to 16×4 -RQAM and 32×2 -RQAM, respectively, for the target ASER of 10^{-4} . Fig. 3(c) compares the ASER curves for different RQAM constellation sizes. It can be seen that, for a given ASER, the required SNR increases with the increase in constellation size. The increased SNR requirement is compensated with the increase in spectral efficiency for higher constellation sizes.

As an important observation, from Fig. 3(b), it can be inferred that, for the same constellation size, SQAM always performs better than non-square RQAM. The reason for sub-optimum performance of non-square RQAM over SQAM is that, in non-square RQAM, the minimum distance between the constellation points is not maximized for a given energy. As a result, non-square RQAM requires slightly more power to achieve a given minimum distance between the constellation points as compared to SQAM.

VII. CONCLUSION

In this work, performance analysis of general order RQAM schemes for dual-hop AF relayed NLOS UV communication system over turbulent channel is conducted. Lower-bound expression for the outage probability is presented and the PDF of end-to-end link is derived. The derived PDF is used to compute the analytical expression of ASER for RQAM. Comparative analysis of different RQAM constellations is performed and it is shown that SQAM always performs better than RQAM with non-square constellation. In future, it will be of interest to study the ASER performance of RQAM for multi-hop AF relays.

VIII. ACKNOWLEDGEMENT

This publication is an outcome of the R&D work undertaken project under the Visvesvaraya PhD Scheme

of Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation.

APPENDIX

On substituting $x = (\ln(\gamma) - \mu)/\sqrt{2}\sigma$ in (20) and (21) and after simplifications, we get

$$I_1(\alpha_1, \mu, \sigma) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \mathcal{Q}\left(\alpha_1 \sqrt{\exp(\sqrt{2}\sigma x + \mu)}\right) e^{-x^2} dx, \quad (24)$$

$$I_2(\alpha_1, \alpha_2, \mu, \sigma) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \mathcal{Q}\left(\alpha_1 \sqrt{\exp(\sqrt{2}\sigma x + \mu)}\right) \mathcal{Q}\left(\alpha_2 \sqrt{\exp(\sqrt{2}\sigma x + \mu)}\right) e^{-x^2} dx. \quad (25)$$

Integrals in (24) and (25) can be evaluated using Gauss-Hermite quadrature integration as

$$I_1(\alpha_1, \mu, \sigma) = \frac{1}{\sqrt{\pi}} \sum_{i=1}^n w_i \mathcal{Q}\left(\sqrt{\alpha_1 \exp(\sqrt{2}\sigma x_i + \mu)}\right), \quad (26)$$

$$I_2(\alpha_1, \alpha_2, \mu, \sigma) = \frac{1}{\sqrt{\pi}} \sum_{i=1}^n \left(w_i \mathcal{Q}\left(\alpha_1 \sqrt{\exp(\sqrt{2}\sigma x_i + \mu)}\right) \mathcal{Q}\left(\alpha_2 \sqrt{\exp(\sqrt{2}\sigma x_i + \mu)}\right) \right), \quad (27)$$

where x_i and w_i are the zeros and weight factors, respectively, of n^{th} order Hermite polynomial. Values of x_i and w_i for different n are provided in [18, Table-25.10].

REFERENCES

- [1] Z. Xu and B. M. Sadler, "Ultraviolet communications: potential and state-of-the-art," *IEEE Commun. Mag.*, vol. 46, no. 5, May 2008.

- [2] Z. Yong, W. Jian, X. Houfei, and L. Jintong, "Non-line-of-sight ultraviolet communication performance in atmospheric turbulence," *China Commun.*, vol. 10, no. 11, pp. 52–57, Nov. 2013.
- [3] M. H. Ardakani, A. R. Heidarpour, and M. Uysal, "Performance analysis of relay-assisted NLOS ultraviolet communications over turbulence channels," *IEEE/OSA J. of Opt. Commun. Net.*, vol. 9, no. 1, pp. 109–118, Jan. 2017.
- [4] M. H. Ardakani and M. Uysal, "Relay-assisted OFDM for ultraviolet communications: performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 607–618, Jan. 2017.
- [5] J. N. Laneman, D. N. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.
- [6] N. Kumar, P. K. Singya, and V. Bhatia, "ASER analysis of hexagonal and rectangular QAM schemes in multiple-relay networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1815–1819, Feb. 2018.
- [7] Q. He, Z. Xu, and M. S. Brian, "Non-line-of-sight serial relayed link for optical wireless communications," in *Military Commun. conf (Milcom)*, Oct. 2010, pp. 1588–1593.
- [8] O. Burdakov, P. Doherty, K. Holmberg, and P.-M. Ols-son, "Optimal placement of UV-based communications relay nodes," *J. Global Optim.*, vol. 48, no. 4, pp. 511–531, Dec. 2010.
- [9] S. Arnon, J. Barry, and G. Karagiannidis, *Advanced optical wireless communication systems*. Cambridge University Press, 2012.
- [10] D. Dixit and P. Sahu, "Performance analysis of rectangular QAM with SC receiver over Nakagami- m fading channels," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1262–1265, Jul. 2014.
- [11] Z. Xu, H. Ding, B. M. Sadler, and G. Chen, "Analytical performance study of solar blind non-line-of-sight ultraviolet short-range communication links," *Optics Lett.*, vol. 33, no. 16, pp. 1860–1862, Aug. 2008.
- [12] Z. Yong, W. Jian, X. Houfei, and L. Jintong, "Non-line-of-sight ultraviolet communication performance in atmospheric turbulence," *China Commun.*, vol. 10, no. 11, pp. 52–57, 2013.
- [13] S. Arya and Y. H. Chung, "Non-line-of-sight ultraviolet communication with receiver diversity in atmospheric turbulence," *IEEE Photonics Technol. Lett.*, vol. 30, no. 10, pp. 895–898, May 2018.
- [14] M. Uysal, J. Li, and M. Yu, "Error rate performance analysis of coded free-space optical links over gamma-gamma atmospheric turbulence channels," *IEEE Trans. on wireless commun.*, vol. 5, no. 6, pp. 1229–1233, Jun. 2006.
- [15] L. C. Andrews, R. L. Phillips, and C. Y. Hopen, *Laser beam scintillation with applications*. SPIE press, 2001, vol. 99.
- [16] G. Chen, Z. Xu, H. Ding, and B. M. Sadler, "Path loss modeling and performance trade-off study for short-range non-line-of-sight ultraviolet communications," *Optics Express*, vol. 17, no. 5, pp. 3929–3940, Mar. 2009.
- [17] P. K. Singya, N. Kumar, and V. Bhatia, "Impact of imperfect CSI on ASER of hexagonal and rectangular QAM for AF relaying network," *IEEE Commun. Lett.*, vol. 22, no. 2, Feb. 2018.
- [18] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions (National Bureau of Standards, Washington, DC, 1972)*, 10th ed., Jun. 1972.

SSK Performance with SWIPT based Dual-Hop AF Relay over Rayleigh Fading

Hemanta Kumar Sahu

School of Electrical Science
Indian Institute of Technology,
Bhubaneswar, India 752050
Email: hs10@iitbbs.ac.in

P. R. Sahu

School of Electrical Science
Indian Institute of Technology,
Bhubaneswar, India 752050
Email: prs@iitbbs.ac.in

Abstract—A cooperative communication system with space shift keying (SSK) modulation and simultaneous wireless information and power transfer (SWIPT) scheme is proposed. SWIPT can eliminate the need of external power supply at the relay whereas SSK modulation scheme reduces inter-channel interference, excludes inter antenna synchronization requirement and the number of radio frequency chains. An upper bound expression for the average bit error probability (ABEP) is obtained with multiple amplify-forward relays and a direct link from source node to destination node. Further, ABEP is analyzed for partial relay selection operation. Numerical and computer simulation results demonstrate performance improvement for SSK modulation combined with SWIPT.

Index Terms—ABEP, Relay selection, Rayleigh fading, SWIPT, SSK Modulation.

I. INTRODUCTION

Energy harvesting can be used in wireless communication networks to recharge battery from an external energy source like solar, wind, and radio frequency (RF) signals. RF based simultaneous wireless information and power transfer (SWIPT) technology can simultaneously convey information and harvest energy in networks [1]- [9]. Hence, SWIPT, if implemented at the energy constrained nodes having restricted access facility (e.g., production plants, battle fields, remote areas under surveillance etc.) can support the energy need of the relays. Some examples of SWIPT include radio-frequency identification networks, wireless body area networks, and wireless sensor networks. Space shift keying (SSK) modulation is known to offer high data rate with relatively low transmission complexity [10]. Therefore, SSK combined with SWIPT can enhance data rate of limited energy networks.

SWIPT architecture was initially proposed for capacity-energy function in [1] and subsequently it has been applied to cooperative communication fields [3], [4], multiple-input-multiple-output (MIMO) [5], and two way relaying [6] networks. In [7]- [9], a practical SWIPT receiver architecture was implemented where the receiver operates either in a time switch (TS) mode or power splitting (PS) mode. Bit error rate (BER) performance of SSK modulation is studied under cooperative communication with AF relay and DF relay over dual hop and multiple hop links over fading channels in [11]- [14]. Recently, in [15], spatial modulation accompanied by SWIPT has been proposed with full duplex two way AF

relaying network and in [16], the authors have analyzed SSK modulation with SWIPT for ideal channel condition. However, analysis of SWIPT scheme combined with SSK modulation over fading channels will be useful for research community.

In this paper, we investigate the power splitting based SWIPT accompanied by SSK modulation in a cooperative communication system with AF relays. The average bit error probability (ABEP) performance of the receiver is investigated for multiple AF relays with a direct link between source and destination and partial selection of single AF relay from multiple AF relays over Rayleigh fading channels. The contributions in this paper are as follows: i) An upper bound for ABEP is derived with multiple energy harvesting AF relays with 2×1 multiple input single output antenna arrangement for SSK modulation with the direct link. ii) A Partial relay selection scheme is proposed and ABEP has been derived. iii) An accurate and simple asymptotic expression for ABEP is obtained for high signal to noise ratio (SNR).

II. SYSTEM AND CHANNEL MODEL

A SSK system model having two transmitting antennas at source (**S**) and one receiving antenna at destination (**D**) with energy harvesting AF relay (**R**) is shown in Fig. 1. The SSK modulation principle can be described as below

- 1) The total information divides into blocks of length ($\log_2 N_t$). Each block is mapped to a transmitting antenna for indexing to transmit data while all other antennas are kept inactive. Let the m bits ($m = \log_2 N_t$) be mapped to a symbol x_j and is transmitted from the j^{th} antenna. In this case, though x_j does not convey information, its location in x does. So, the vector x_j specifies [10]

$$x_j \equiv [0, 0 \dots, \underbrace{1}_{j^{th} \text{ position}}, 0, 0 \dots]^T$$

- 2) The non ideal transmit antenna estimation is solved at the receiver end using N_t -hypothesis detection problem. However, the transmitted signal is denoted by $S_j(\cdot)$, for $j = 1, 2, \dots, N_t$. It is presented due to m_j is transmitted message, the analog signal $S_j(\cdot)$ is emitted by the j^{th} transmit antenna while other antennas will remain inactive [10].

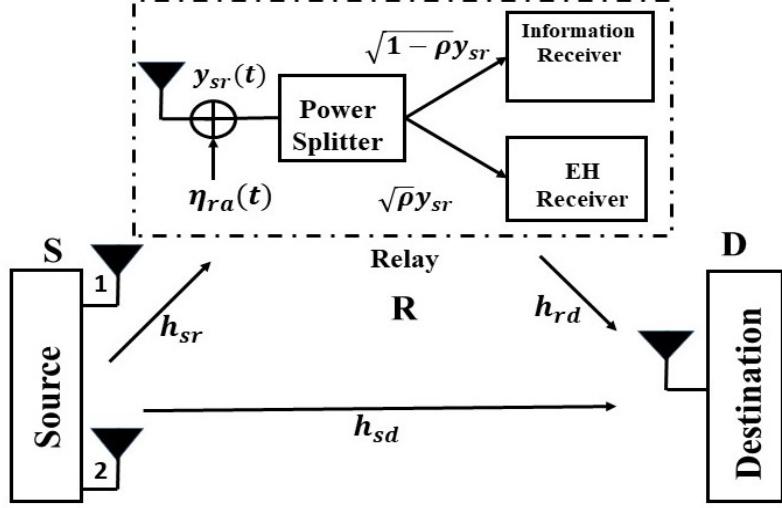


Fig. 1. SSK system model with Energy harvesting AF relay.

Therefore, an expression for the signal received at the relay input can be written as [11]

$$x(t) = \sqrt{E_S}g_l + n_1(t), \quad (1)$$

where g_l , with $l = 1, 2$ is a Rayleigh distributed fading coefficient between the transmitting antenna l and **R**, E_S is energy transmitted from **S** and $n_1(t)$ is a Gaussian noise with zero mean and variance N_0 . The relay **R** divides the received signal power, with a power splitting ratio of $\rho : 1 - \rho$, that utilizes $\sqrt{\rho}$ portion of the received power for energy harvesting and the rest $\sqrt{1 - \rho}$ for the information extraction. The energy harvested at the AF relay node in the first time slot can be written as [2]

$$E_R = \eta E_S |g_l|^2 \rho, \quad (2)$$

where $0 < \eta < 1$ is the energy conversion efficiency, that depends on the energy harvesting circuit rectification process. The signal received at the k^{th} energy harvesting relay information processing unit can be written as [2]

$$y_{R_k D}(t) = \sqrt{1 - \rho} \sqrt{E_S} g_l + \sqrt{1 - \rho} n_1(t) + n_2(t), \quad (3)$$

where $n_2(t)$ is a baseband additive white Gaussian noise with $\mathcal{N}(0, N_0)$ and $k=1, \dots, L$ number of realys. In the second time slot, the k^{th} relay amplifies and forward the information signal using the harvested energy. Thus, the received signal from relay at **D** can be expressed as

$$y_{R_k D}(t) = \beta h \{ \sqrt{1 - \rho} \sqrt{E_S} h_l + \sqrt{1 - \rho} n_1(t) + n_2(t) \} + n_d(t), \quad (4)$$

where h is a Rayleigh fading coefficient between the **R** and the **D**, β is a fixed amplification factor and $n_d(t)$ is an AWGN sample. After normalizing the noise terms ($\beta h \sqrt{1 - \rho} n_1(t) + \beta h n_2(t) + n_d(t)$), the received signal in (4) can be expressed

as

$$y_{R_k D}(t) = \sqrt{H} h g_l + \tilde{w}, \quad (5)$$

where $H = \frac{\beta^2(1-\rho)E_S}{(2-\rho)\beta^2|h|^2+1}$ and \tilde{w} is a Gaussian random variable with $\mathcal{N}(0, N_0)$. Similarly, the signal received through direct link between **S** to **D** can be written as

$$y_{SD}(t) = \sqrt{E_S} f_l + n_3(t), \quad (6)$$

where f_l is a non line of sight Rayleigh fading channel coefficient between **S** to **D**.

III. ANALYSIS OF ABEP

Optimal Detection: Since the channel inputs are assumed equally likely, the maximum likelihood (ML) optimal detector can be expressed as

$$\tilde{l} = \arg \max \{D_m\}, \quad (7)$$

where D_m denotes the decision metric, defined as [17]

$$D_m = \Re\{y_{R_k D}(t) \times \sqrt{H} h^* g_m^*\} - \frac{1}{2} H |h|^2 |g_m|^2 + \Re\{y_{SD}(t) \times \sqrt{E_S} f_m^*\} - \frac{1}{2} |f_m|^2, \quad (8)$$

where $|\cdot|$, $(\cdot)^*$, and $\Re\{\cdot\}$ denote the absolute value, complex conjugate, and real part, respectively. The decision metrics D_m in (8) can be rewritten by activating the l^{th} transmit antenna as

$$D_{m|l=m} = \frac{H}{2} |h|^2 |g_l|^2 + \sqrt{H} \Re\{h^* g_l^* \tilde{w}\} + \frac{E_S}{2} |f_l|^2 + \sqrt{E_S} |f_l| n_3(t) \\ D_{m|l \neq m} = H |h|^2 \Re\{g_l g_m^*\} - \frac{H}{2} |h| |g_m|^2 + \sqrt{H} \Re\{h^* g_m^* \tilde{w}\} + \frac{E_S}{2} \Re\{f_l f_m^*\} - \frac{E_S}{2} |f_l|^2 + \sqrt{E_S} |f_l| n_3(t). \quad (9)$$

A. Multiple Relays with direct link

Assuming the detection of incorrect transmit-antenna index, the conditional probability of error $P_e(g_1, g_2, h, f_1, f_2)$, can be expressed by considering two transmitting antennas, a single relay and the direct link as

$$P_e(g_1, g_2, h, f_1, f_2) = \frac{1}{2} Pr(D_1|l=1 < D_2|l=1) + \frac{1}{2} Pr(D_2|l=2 < D_1|l=2). \quad (10)$$

Substituting (9) in (10), and after some algebraic manipulations, the $P_e(g_1, g_2, h, f_1, f_2)|_{l=1}$ can be written as

$$P_e(g_1, g_2, h, f_1, f_2)|_{l=1} = Pr\left(H|h|^2|g_2 - g_1|^2 + \frac{E_S}{2}|f_2 - f_1|^2 < \hat{w}\right), \quad (11)$$

where $\hat{w} = \sqrt{E_S} \Re\{n_3(f_2^* - f_1^*)\} + \sqrt{H} \Re\{\tilde{w}h^*(g_2^* - g_1^*)\}$, is a Gaussian distributed RV with zero-mean and variance $2N_0(E_S|f_2 - f_1|^2 + H|g_2 - g_1|^2)$. Thus, the general conditional error probability either considering $l = 1$ or $l = 2$ with L number of relays can be written as

$$P_e(g_1, g_2, h, f_1, f_2) = Q\left(\sqrt{\gamma_{sd} + \sum_{i=1}^L Z_i}\right), \quad (12)$$

where $\gamma_{sd} = \frac{P_S}{2}|f_2 - f_1|^2$, $Z_i = \frac{(1-\rho)\alpha_R\alpha_S}{(2-\rho)\alpha_R+D}$, $\alpha_R = P_R|h|^2$, $\alpha_S = \frac{P_S|g_2 - g_1|^2}{2}$, $P_R = \frac{E_R}{N_0}$, $P_S = \frac{E_S}{N_0}$, $D = \frac{E_R}{\beta^2 N_0}$ and $Q(\cdot)$ represents the Gaussian Q-function. Note that α_S and γ_{sd} are exponential distributed random variable with CDF of α_S is given as $F_{\alpha_S}(x) = 1 - \exp\left(\frac{-x}{P_S\Omega_1}\right)$, where $\Omega_1 = d_1^{-\alpha}$ is the average channel gain, d_1 is the distance between **S** and **R**, and $\alpha \in [2,5]$ is the path-loss exponent. Similarly, the CDF of γ_{sd} is $F_{\gamma_{sd}}(x) = 1 - \exp\left(\frac{-x}{P_S\Omega_3}\right)$ with $\Omega_3 = d_3^{-\alpha}$, d_3 is the distance between **S** and **D**. Now, substituting E_R from (2) the random variable α_R can be given as $\alpha_R = \beta|g_1|^2|h|^2$, where $\beta = \eta P_S \rho$. To find the probability density function (PDF) of α_R , we first need to find the PDF of $\Upsilon = |g_1|^2|h|^2$, where the PDF of $|h|^2$ is $f_{|h|^2}(x) = \frac{1}{\Omega_2} \exp\left(\frac{-x}{\Omega_2}\right)$ with $\Omega_2 = d_2^{-\alpha}$, d_2 is the distance between **R** and **D**. Since, h and g_1 are statistically independent, the PDF of Υ can be obtained as

$$f_{\Upsilon}(\lambda) = \int_0^{\infty} \frac{1}{x} f_{|h|^2}(x) f_{|g_1|^2}\left(\frac{\lambda}{x}\right) dx = \int_0^{\infty} \frac{1}{x\Omega_1\Omega_2} \exp\left(-\frac{x}{\Omega_1} - \frac{\lambda}{\Omega_2 x}\right) dx. \quad (13)$$

The above integration can be evaluated using the identity [18, (3.471-9)] as

$$f_{\Upsilon}(\lambda) = \frac{2}{\Omega_1\Omega_2} K_0\left(2\sqrt{\frac{\lambda}{\Omega_1\Omega_2}}\right), \quad (14)$$

where $K_v(\cdot)$ is the v -th order Bessel function of second kind [18]. Thus, using the principle of transformation of random

variables the PDF of α_R can be obtained as

$$f_{\alpha_R}(\lambda) = \frac{2}{\beta\Omega_1\Omega_2} K_0\left(2\sqrt{\frac{\lambda}{\beta\Omega_1\Omega_2}}\right). \quad (15)$$

The ABEP can be derived from (12), by exploiting the moment-generation function (MGF)-based approach. The MGF of γ_{sd} is [17]

$$\mathcal{M}_{\gamma_{sd}}(s) = \frac{1}{1+sP_S\Omega_3}. \quad (16)$$

Similarly, to get the MGF of Z_i , first we need to find the CDF $F_{Z_i}(\gamma)$. Hence,

$$F_{Z_i}(\gamma) = Pr(Z_i < \gamma) = Pr\left(\frac{(1-\rho)\alpha_R\alpha_S}{(2-\rho)\alpha_R+D} < \gamma\right) = \int_0^{\infty} Pr\left[\alpha_S < \frac{x((2-\rho)\alpha_R+D)}{(1-\rho)\alpha_R} \middle| \alpha_R\right] f_{\alpha_R}(\alpha_R) d\alpha_R. \quad (17)$$

Now substituting the CDF F_{α_S} and PDF f_{α_R} in (17), the integral can be rewritten as

$$F_{Z_i}(\gamma) = \frac{2}{\beta\Omega_1\Omega_2} \int_0^{\infty} \left[1 - e^{-\frac{\gamma}{P_S\Omega_1}\left(\frac{2-\rho}{1-\rho} + \frac{D}{(1-\rho)\alpha_R}\right)}\right] \times K_0\left(2\sqrt{\frac{\alpha_R}{\beta\Omega_1\Omega_2}}\right) d\alpha_R. \quad (18)$$

To find MGF, it requires PDF of Z_i , which can be obtained by differentiating (18) with respect to γ , $\forall i$, as

$$f_{Z_i}(\gamma) = \frac{2}{P_S\beta\Omega_1^2\Omega_2} \left[\left(\frac{2-\rho}{1-\rho}\right) \int_0^{\infty} e^{-\frac{\gamma}{P_S\Omega_1}\left(\frac{2-\rho}{1-\rho} + \frac{D}{(1-\rho)\alpha_R}\right)} \times K_0\left(2\sqrt{\frac{\alpha_R}{\beta\Omega_1\Omega_2}}\right) d\alpha_R + \frac{D}{1-\rho} \int_0^{\infty} \alpha_R^{-1} e^{-\frac{\gamma}{P_S\Omega_1}\left(\frac{2-\rho}{1-\rho} + \frac{D}{(1-\rho)\alpha_R}\right)} K_0\left(2\sqrt{\frac{\alpha_R}{\beta\Omega_1\Omega_2}}\right) d\alpha_R \right]. \quad (19)$$

To solve the above integral in (19), we express the exponential and Bessel function terms by Meijer's G-function [19]. Thus (19) can be rewritten as

$$f_{Z_i}(\gamma) = \frac{1}{P_S\beta\Omega_1^2\Omega_2} e^{-\mu\gamma} (\psi\gamma) \left[\frac{2-\rho}{1-\rho} \int_0^{\infty} G_{1\ 0}^{0\ 1}(\alpha_R|.) \times G_{0\ 2}^{2\ 0}(\Theta\gamma\alpha_R|0\ 0) d\alpha_R + \frac{D}{1-\rho} \int_0^{\infty} \alpha_R^{-1} G_{1\ 0}^{0\ 1}(\alpha_R|.) \times G_{0\ 2}^{2\ 0}(\Theta\gamma\alpha_R|0\ 0) d\alpha_R \right]. \quad (20)$$

where $\mu = \frac{1}{P_S\Omega_1}\left(\frac{2-\rho}{1-\rho}\right)$, $\psi = \frac{D}{P_S\Omega_1(1-\rho)}$, and $\Theta = \frac{D}{P_S\beta\Omega_1^2\Omega_2(1-\rho)}$. With the help of [20], the PDF $f_{Z_i}(\gamma)$ is obtained in closed form as

$$f_{Z_i}(\gamma) = \frac{1}{\beta P_S\Omega_1^2\Omega_2} e^{-\mu\gamma} (\psi\gamma) \left[\frac{2-\rho}{1-\rho} G_{0\ 3}^{3\ 0}(\Theta\gamma|0\ 0\ 0) + \frac{D}{1-\rho} G_{0\ 3}^{3\ 0}(\Theta\gamma|0\ 0\ 0) \right]. \quad (21)$$

However, for multiple number (L) of relays, we need to find the PDF of δ where $\delta = \sum_{i=1}^L Z_i$ that involves L convolutions

and hence is not straight forward to obtain. Hence, we resort to MGF technique. The MGF of Z_i can be given as

$$\mathcal{M}_{Z_i}(s) = \int_0^\infty e^{-s\gamma} f_{Z_i}(\gamma) d\gamma. \quad (22)$$

Substituting $f_{Z_i}(\gamma)$ from (21), in (22) and using identity [20], the MGF can be obtained as in (23) shown in top of the next page. Hence, the ABEP can be evaluated as [12]

$$ABEP = \frac{1}{\pi} \int_0^{\pi/2} \mathcal{M}_{sd} \left(\frac{1}{2 \sin^2(\theta)} \right) \prod_{i=1}^L \mathcal{M}_{Z_i} \left(\frac{1}{2 \sin^2(\theta)} \right) d\theta. \quad (24)$$

Thus, substituting (16) and (23) into (24), an exact expression for the ABEP in a finite single integral can be obtained. To avoid numerical integration, (24) can be upper bounded as [12]

$$ABEP < \frac{1}{\pi} \mathcal{M}_{sd} \left(\frac{1}{2} \right) \prod_{i=1}^L \mathcal{M}_{Z_i} \left(\frac{1}{2} \right). \quad (25)$$

B. Asymptotic Analysis

To get the better in the system performance, we present asymptotic analysis for high value of P_S . Further the asymptotic expression of $\mathcal{M}_{Z_i}(s)$ can be simply obtained by using the Meijer's G-function expansion in (23). Assuming $\omega = \frac{D}{\beta \Omega_1 \Omega_2 (s P_S \Omega_1 (1-\rho) + 2(2-\rho))}$ and using the condition $\omega \rightarrow 0$, $p > q$, and $\lim_{\omega \rightarrow 0^+} p F_q(a, b, \omega) = 1$ the expression in (23) can be approximated as [21, (18)]

$$\begin{aligned} \mathcal{M}_{Z_i}^{asym}(s) &= \frac{1}{\beta P_S \Omega_1^2 \Omega_2} \left[\frac{D(2-\rho)}{P_S \Omega_1^2 (1-\rho)^2} \frac{1}{(s+\mu)^2} \sum_{k=1}^3 (\omega)^{\eta_{2,k}} \right. \\ &\quad \times \left. \frac{\Gamma(1+\eta_{2,k} - \eta_{1,1}) \prod_{j=1, j \neq k}^3 \Gamma(\eta_{2,j} - \eta_{2,k})}{\Gamma(\eta_{1,2} - \eta_{2,k}) \prod_{j=2}^3 \Gamma(1+\eta_{2,k} - \eta_{2,j})} \right], \end{aligned} \quad (26)$$

where $\eta_{a,b}$ denotes the b^{th} term of η_a with $\eta_1 = \{-1\}$ and $\eta_2 = \{0, 0, -1\}$ and the second Meijer's G-function value is zero in (23) for high value of P_S .

C. Partial Relay Selection with direct link

In this technique, the relay that has the best link between S and R_k is selected. So, the Euclidean distances between S to R_k decides the best relay selection. For partial relay selection α_S can be written as $\alpha_S^{PRS} = \max_{1 \leq k \leq L} \alpha_{Sk}$. So, the CDF of

α_S can be expressed as $F_{\alpha_S^{PRS}}(\gamma) = \left(1 - e^{-\frac{\gamma}{P_S \Omega_1}}\right)^L$. Using partial relay selection with direct link, the conditional error probability can be rewritten as

$$P_e(PRS) = Q \left(\sqrt{\gamma_{sd} + Z^{PRS}} \right). \quad (27)$$

Now substituting the CDF $F_{\alpha_S^{PRS}}(\gamma)$ in (17), the expression for $F_{Z^{PRS}}(\gamma)$ can be written as

$$\begin{aligned} F_{Z^{PRS}}(\gamma) &= \frac{2}{\beta \Omega_1 \Omega_2} \int_0^\infty \left[1 + \sum_{m=0}^L \binom{L}{m} (-1)^m \right. \\ &\quad \times \left. e^{-\frac{m\gamma}{P_S \Omega_1} \left(\frac{2-\rho}{1-\rho} + \frac{D}{(1-\rho)\alpha_R} \right)} \right] K_0 \left(2 \sqrt{\frac{\alpha_R}{\beta \Omega_1 \Omega_2}} \right) d\alpha_R. \end{aligned} \quad (28)$$

For the MGF based approach, the PDF ' $f_{Z^{PRS}}$ ' needs to be derived from (28) by differentiation. So, the expression of $f_{Z^{PRS}}$ can be derived as the same procedure in (20) and expressed as

$$\begin{aligned} f_{Z^{PRS}}(\gamma) &= \frac{1}{P_S \beta \Omega_1^2 \Omega_2} \sum_{m=1}^L (-1)^m e^{-m\mu\gamma} (m\psi\gamma) \left[\frac{2-\rho}{1-\rho} \right. \\ &\quad \times \int_0^\infty G_{1,0}^{0,1}(\alpha_R | \cdot) G_{0,2}^{2,0}(\Theta m\gamma\alpha_R | \cdot \cdot) d\alpha_R \\ &\quad \left. + \frac{D}{1-\rho} \int_0^\infty \alpha_R^{-1} G_{1,0}^{0,1}(\alpha_R | \cdot) G_{0,2}^{2,0}(\Theta m\gamma\alpha_R | \cdot \cdot) d\alpha_R \right]. \end{aligned} \quad (29)$$

The above integration in (29), can be evaluated as [20]

$$\begin{aligned} f_{Z^{PRS}}(\gamma) &= \frac{1}{\beta P_S \Omega_1^2 \Omega_2} \sum_{m=1}^L (-1)^m e^{-m\mu\gamma} (\psi\gamma) \left[\frac{2-\rho}{1-\rho} \right. \\ &\quad \times G_{0,3}^{3,0}(\Theta m\gamma | \cdot \cdot \cdot \cdot \cdot \cdot \cdot) + \left. \frac{D}{1-\rho} G_{0,3}^{3,0}(\Theta m\gamma | \cdot \cdot \cdot \cdot \cdot \cdot \cdot) \right]. \end{aligned} \quad (30)$$

The MGF of Z^{PRS} can be derived by substituting $f_{Z^{PRS}}(\gamma)$ in (22) and using the identity [18, (7.811-3)] as shown in (31) in the next page. The ABEP can be evaluated by substituting (16) and (31) in (24) and to avoid numerical integration the upper bound approximation (25) can be applied.

IV. NUMERICAL AND SIMULATION RESULTS

Numerically evaluated results and computer simulation results are presented. In the evaluation, we use the path loss exponent, $\alpha = 2$ and the energy harvesting efficiency, $\eta = 1$. Setting η less than one, will degrade the system performance, as it would result less amount of energy harvesting at the relays. For the convenience of numerical and subsequent power allocation purpose, we assume $\sigma_g^2 = \sigma_h^2 = \sigma_f^2 = 1$, $N_0 = 1$, and $P_T = \frac{E_s}{N_0} + \frac{E_R}{N_0}$.

In Fig. 2, ABEP vs. ρ is plotted for different values of P_T and for different number of relays with direct link. It can be observed from the curves that combining SSK modulation and energy harvesting AF relay, the ABEP performance improves as ρ increases from 0.1 to a value ρ_m and then it degrades with $\rho > \rho_m$. For example the value of $\rho_m = 0.3$ for $P_T = 10$ dB. The above ABEP variation is due to the fact that with $\rho < \rho_m$, low power is accessible for harvesting energy and hence the energy available at AF relay node for signal transmission is lower, hence the ABEP is high. However, for $\rho > \rho_m$, the amount of power utilization in energy harvesting circuit is more and

$$\begin{aligned} \mathcal{M}_{Z_i}(s) = & \frac{1}{\beta P_S \Omega_1^2 \Omega_2} \left[\frac{D(2-\rho)}{P_S \Omega_1^2 (1-\rho)^2} \frac{1}{(s+\mu)^2} G_{1 \cdot 3}^{3 \cdot 1} \left(\frac{D}{\beta \Omega_1 \Omega_2 (s P_S \Omega_1 (1-\rho) + 2(2-\rho))} \middle| \begin{array}{ccc} -1 & \cdot & \cdot \\ 0 & 0 & -1 \end{array} \right) \right. \\ & \left. + \left(\frac{D}{1-\rho} \right) \frac{1}{s+\mu} G_{1 \cdot 3}^{3 \cdot 1} \left(\frac{D}{\beta \Omega_1 \Omega_2 (s P_S \Omega_1 (1-\rho) + 2(2-\rho))} \middle| \begin{array}{ccc} 0 & \cdot & \cdot \\ 0 & 0 & 0 \end{array} \right) \right]. \end{aligned} \quad (23)$$

$$\begin{aligned} \mathcal{M}_{ZPRS}(s) = & \frac{1}{\beta P_S \Omega_1^2 \Omega_2} \sum_{m=0}^L (-1)^m \left[\frac{D(2-\rho)}{P_S \Omega_1^2 (1-\rho)^2} \frac{m}{(s+\mu)^2} G_{1 \cdot 3}^{3 \cdot 1} \left(\frac{mD}{\beta \Omega_1 \Omega_2 (s P_S \Omega_1 (1-\rho) + 2(2-\rho))} \middle| \begin{array}{ccc} -1 & \cdot & \cdot \\ 0 & 0 & -1 \end{array} \right) \right. \\ & \left. + \left(\frac{D}{1-\rho} \right) \frac{m}{s+\mu} G_{1 \cdot 3}^{3 \cdot 1} \left(\frac{mD}{\beta \Omega_1 \Omega_2 (s P_S \Omega_1 (1-\rho) + 2(2-\rho))} \middle| \begin{array}{ccc} 0 & \cdot & \cdot \\ 0 & 0 & 0 \end{array} \right) \right]. \end{aligned} \quad (31)$$

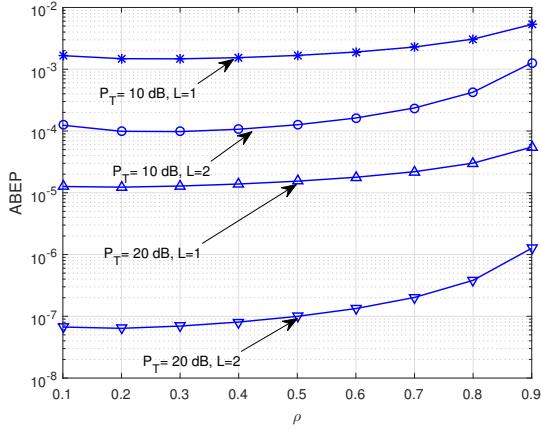


Fig. 2. ABEP versus ρ with varying P_T .

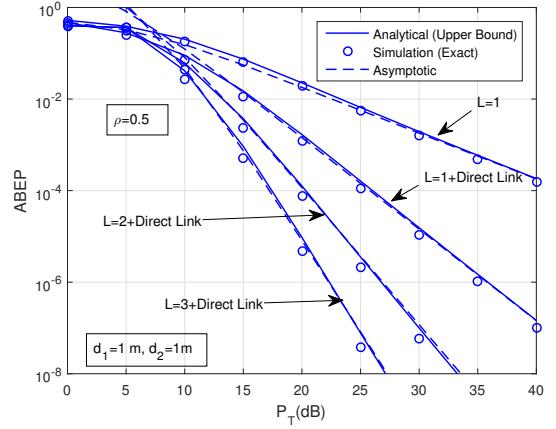


Fig. 3. ABEP of SSK with multiple relays.

the power recovery of the signal from the transmitter is low resulting for the transmission of a noisy signal from **R** to **D**. In Fig. 3, increasing the number of relays with direct link from **S** to **D**, the ABEP performance improves as shown. In the figure, the curves represent the performance of the ABEP vs. P_T , for varying number of energy harvesting AF relays, $L=1,2,3$. with direct link. It can be noticed that as L increases, the ABEP improves significantly due to the combined effects of cooperative diversity and harvesting energy available at the relays. Again a comparison has been made between with direct link and without direct link. It can be observed that $L=1$ with direct link has 8 dB improvement of P_T in compare with only relay to achieve an ABEP of 10^{-2} . Also from Fig. 3, it is clear that the asymptotic results match with the simulated and numerical results for high values of P_T . Fig. 4 plots the ABEP vs. P_T with varying distances d_1 and d_2 , where d_1 and d_2 are the distance between **S** to **R** and **R** to **D**, respectively. As d_1 and d_2 varies the value of d_3 changes. It can be observed from the figure that the ABEP degrades as d_1 increases, and it is not affected as d_2 increases. It is because as d_1 increases the received power at the energy harvesting relays decreases causing the decrease in transmitted power and thus deteriorates

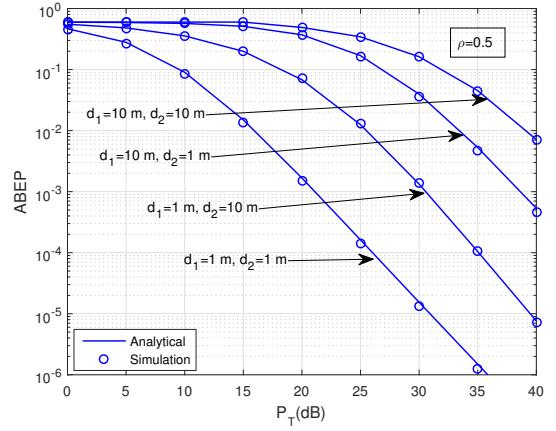


Fig. 4. ABEP of SSK with varying distance.

the ABEP. For example, decreasing on d_1 from 10 meter to 1 meter ($d_2=1$ meter fixed), to achieve an ABEP of 10^{-3} , the reduction in P_T is about 18 dB. Similarly, in Fig. 5 we plot ABEP vs. d_2 by assuming $d_1=1$ meter (fixed) and $P_T=30$ dB. The curves show the variation of the number of relays with a

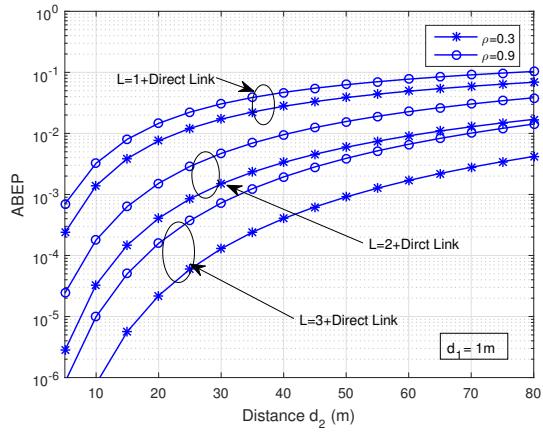


Fig. 5. ABEP versus distance for varying number of relays.

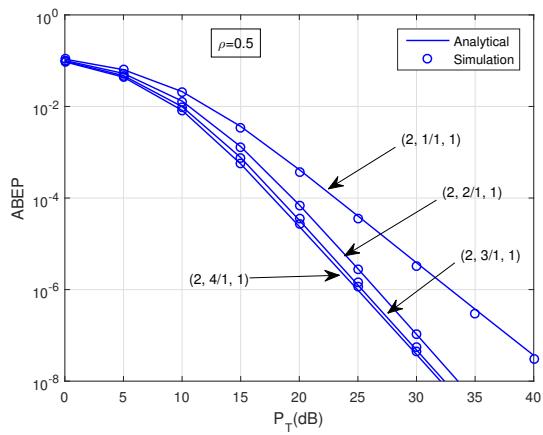


Fig. 6. ABEP of SSK with partial relay selection.

direct link and ρ . It can be observed that better performance can be achieved with the increase number of relays with direct link or with the decrease in power splitting ratio ρ . It is because as more relays more energy can be harvested so that the distance between relay to the receiver can be increased.

To reduce the hardware complexity needed for time synchronization between multiple relays, the performance of partial relay selection is shown in Fig. 6. For simplicity, we use $(N_t, L/1, 1)$ notation where N_t is the number of transmit antennas, $L/1$ is the partial selection of one relay out of L number of relays, and the last one is the receiving antenna. In Fig. 6, it can be observed that by increasing the number of relays with partial relay selection technique having a constant number of transmitting antennas and receiving antenna, better performance is achieved.

V. CONCLUSION

ABEP performance of cooperative communication AF relaying with SSK modulation and SWIPT is analyzed over Rayleigh fading channel. Results demonstrate ABEP with

the increase in power splitting factor to a certain value and then starts descending. ABEP can also be improved using multiple numbers of energy harvesting relays or with relay selection from multiple relays. The derived expressions have been numerically evaluated and verified against computer simulation results. The work can be further extended for practical nonlinear energy harvesting circuit model.

REFERENCES

- [1] L. Liu, R. Zhang, and K.C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 288-300, Jan. 2013.
- [2] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer: Architecture design and rate-energy tradeoff," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4754-4767, Nov. 2013.
- [3] A. A. Nasir, X. Zhou, S. Durrani, and R. A. Kennedy, "Relaying protocols for wireless energy harvesting and information processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3622-3636, Jul. 2013.
- [4] Xiangli Liu, Zan Li, and Ce Wang, "Secure decode-and-forward relay SWIPT systems with power splitting scheme," *IEEE Trans. Veh. Techno.*, DOI 10.1109/TVT.2018.2833446.
- [5] E. Boshkovska, D. W. K. Ng, N. Zlatanov, A. Koelpin, and R. Schober, "Robust resource allocation for MIMO wireless powered communication networks based on a non-linear EH model," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 1984-1999, May 2017.
- [6] Y. Ye, et al., "Dynamic asymmetric power splitting scheme for SWIPT-based two-way multiplicative AF relaying," *IEEE Signal Process Lett.*, vol. 25, no. 7, pp. 1014-1018, July. 2018.
- [7] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer in multiuser OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 2282-2294, Apr. 2014.
- [8] X. Lu, P. Wang, D. Niyato, and Z. Han, "Resource allocation in wireless networks with RF energy harvesting and transfer," *IEEE Netw.*, vol. 23, no. 5, pp. 620-624, May 2016.
- [9] L. Mohjazi, S. Muhamadat, and M. Dianati, "Performance analysis of differential modulation in SWIPT cooperative networks," *IEEE Signal Process. Lett.*, vol. 29, no. 6, pp. 68-75, Nov. 2015.
- [10] J. Jeganathan, A. Ghrayeb, L. Szczecinski, and A. Ceron, "Space shift keying modulation for MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3692-3703, July 2009.
- [11] R. Mesleh, S. Ikki, and M. Alwakeel, "Performance analysis of space shift keying with amplify and forward relaying," *IEEE Commun. Lett.*, vol. 15, no. 12, pp. 1350-1352, Dec. 2011.
- [12] R. Mesleh, S. Ikki, H. Aggoune, A. Mansour, "Performance analysis of space shift keying (SSK) modulation with multiple cooperative relays," *Eurasip J. Adv. Signal Process.*, vol. no. 1, pp. 201-210, Sep. 2012.
- [13] M. Wen, X. Cheng, H. V. Poor, and B. Jiao, "Use of SSK modulation in two-way amplify-and-forward relaying," *IEEE Trans. Veh. Technol.*, vol. 63, no. 3, pp. 1498-1504, Mar. 2014.
- [14] P. Som and A. Chockalingam, "Performance analysis of space-shift keying in decode-and-forward multihop MIMO networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 1, pp. 132-146, Jan. 2015.
- [15] A. Koc, I. Altunbas, and E. Basar, "Two-way full-duplex spatial modulation systems with wireless powered AF relaying," *IEEE wireless Commun. Lett.*, vol. 7 no. 3 pp. 444-447, Jun. 2018.
- [16] H. K. Sahu and P. R. Sahu, "Simultaneous wireless information and power transfer with SSK modulation over Rayleigh fading," *Internet Technology Letters*, pp. 1-6, 2018. <https://doi.org/10.1002/itl2.67>
- [17] M. K. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*, 2nd ed. New York: Wiley, 2005.
- [18] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th edition. Academic Press, 2000.
- [19] A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev, *Integrals, and Series : More Special Functions*, Gordon and Breach Sci. Publ., New York, 1990, Vol. 3.
- [20] Available: <http://functions.wolfram.com/> / HypergeometricFunctions / MeijerG/21/02/02/.
- [21] V. S. Adamchik and O. I. Marichev, "The algorithm for calculating integrals of hypergeometric type function and its realization in reduce system," in *Proc. Int. Conf. Symbolic Algebraic Comput.*, 1990, pp. 212-224.

Interference Violation Probability Constrained Underlay Cognitive Massive MIMO Network Under Imperfect Channel Knowledge

Rama Gupta, Salil Kashyap, and E. Venkata Pothan

Abstract—We investigate the use of massive number of antennas at the cognitive base station (BS) in reducing interference caused to primary users (PUs) under imperfect channel knowledge without deteriorating the data rate provided to the cognitive user (CU). To this end, we develop a simple back-off factor based power adaptation policy for the cognitive BS which ensures that its transmissions do not violate the interference violation probability constraint at the PUs. We derive a new lower bound on the complement of the interference violation probability and also deduce a lower bound on the achievable rate of the CU when the cognitive BS has an imperfect estimate of its channels to the PUs and the CU. Through our analytical and numerical results, we quantify that the interference violation probability at the PUs can be reduced while providing a fixed data rate to the CU by deploying more number of cognitive BS antennas. Furthermore, if the number of PUs in the network increase, we show that the interference violation probability at the PUs and the data rate of the CU can be maintained at the same level by increasing the number of antennas at the cognitive BS.

Index Terms—Underlay cognitive radio, massive MIMO, imperfect CSI, achievable rate, array gain

I. INTRODUCTION

Cognitive radio (CR) promises huge improvements in spectrum utilization efficiency for wireless systems and is, therefore, a potential next generation non-orthogonal multiple access technology [1]. In it, there are two classes of users, namely the primary users (PUs), who hold the license to communicate in the allotted spectrum and the cognitive users (CUs) who are the unlicensed users and look for opportunities to communicate over the same spectrum without adversely affecting the performance of the PUs. In this work, we focus on underlay CRs, in which cognitive transmissions can occur concurrently with primary transmissions. However, on the downlink, the interference generated by the cognitive base station (BS) at the PUs must be below an acceptable threshold [2]. This interference constraint together with the accuracy of the channel state information (CSI) from the cognitive BS to the PUs restrict the power with which a cognitive BS can transmit and therefore, limits the data rate that can be offered to the CUs.

Massive multiple input multiple output (MIMO) is a potential fifth generation (5G) cellular technology in which

S. Kashyap and E. V. Pothan are with IIT Guwahati, India. Rama Gupta was with IIT Guwahati during the course of this work. She is currently with Intel, India.

Emails: {rama8gupta, salilkashyap}@gmail.com, potha176102102@iitg.ac.in

hundreds of antennas at the BS can concurrently serve tens of users through spatial multiplexing over the same time-frequency resource [3]. A massive MIMO BS can discriminate users more accurately with narrower and focussed beams and has the ability to enhance spatial division and multiplexing gain [4]. It is being considered as a promising solution for maximizing spectrum reuse [5]. It, therefore, holds enormous potential in improving the data rate in CR systems, where interference caused to the PUs and the accuracy of the CSI to the PUs are of prime concern.

In this paper, we analyze the downlink of such a unified underlay cognitive massive MIMO network that is subject to an interference violation probability constraint, to address the following question: Can we exploit the huge array gain and the high spatial resolution offered by massive number of antennas at the cognitive BS to help reduce the interference violation probability at the PUs 1) without incurring a degradation in the rate offered to the CU, and 2) when the cognitive BS has an imperfect knowledge of its channels to the PUs and the CU? To this end, we make the following specific contributions.

Our Contributions:

- 1) We propose a simple back-off factor based power adaptation policy for the cognitive BS under imperfect CSI, where the back-off factor is chosen such that a target interference violation probability is satisfied at the PUs.
- 2) We derive a new closed-form lower bound for the complement of the interference violation probability at the PUs with imperfect CSI.
- 3) We then develop a lower bound on the achievable rate of the CU when the cognitive BS has an imperfect knowledge of its channels to the PUs and the CU, and transmits based on the proposed policy to meet the target violation probability.
- 4) We present extensive numerical results to obtain insights into the impact of different system parameters such as back-off factor, channel estimation error, the number of PUs and the number of cognitive BS antennas on interference violation probability and the achievable rate.

Related Literature: A multiuser massive MIMO primary network and a multiple-input single-output underlay cognitive network was considered in [6] and the data rates were derived for CU under a peak interference constraint. An asymptotic analysis was also presented both with perfect and imperfect CSI. The authors in [7] considered uplink of an underlay

cognitive massive MIMO network and optimized energy efficiency while ensuring fairness among CUs. The authors in [8] considered an underlay cognitive massive MIMO network and optimized the number of CUs that can be served on the downlink under power, rate, and a peak interference constraint and with imperfect CSI. While the authors in [9] considered a multi-cell massive MIMO primary network and a single cell massive MIMO cognitive network and studied power allocation with pilot contamination to maximize the downlink sum rate. In [10], the authors considered uplink of a single-cell underlay cognitive massive MIMO network under imperfect CSI and investigated joint pilot and data power allocation while ensuring max-min fairness to users.

In contrast to the existing literature, we focus on evaluating the performance of an underlay cognitive massive MIMO BS that is subject to a practically motivated interference violation probability constraint and transmits based on the proposed back-off factor based power adaptation policy to counter the adverse effects of imperfect CSI.

Notation: A circular symmetric complex Gaussian random vector \mathbf{x} with zero mean and covariance matrix \mathbf{A} is denoted by $\mathbf{x} \sim \mathcal{CN}(\mathbf{0}, \mathbf{A})$. We denote conjugate transpose operator by $(\cdot)^\dagger$, transpose by $(\cdot)^T$, the Euclidean norm by $\|\cdot\|$. Furthermore, $\mathbb{E}[\cdot]$ denotes the expectation operator, $\mathbf{0}$ denotes the zero vector, \mathbf{I}_N denotes the $N \times N$ identity matrix, $\text{tr}(\cdot)$ denotes the trace of a matrix, $f_X(x)$ denotes the probability density function (pdf) of a random variable (rv) X and $F_X(x)$ denotes its cumulative distribution function (CDF), $f_{X,Y}(x,y)$ denotes the joint pdf of two rvs X and Y , $\binom{n}{k}$ denotes the $(n$ choose k) operation and the joint probability of events A_1, A_2, \dots, A_n is denoted by $\Pr(A_1, A_2, \dots, A_n)$.

II. SYSTEM MODEL

We consider the downlink of an underlay cognitive massive MIMO network as shown in Figure 1. The system comprises of a primary network that owns the license of the spectrum and a cognitive network that opportunistically accesses the spectrum subject to a constraint on the interference that it can cause to the PUs. The primary network consists of a primary base station (PBS) equipped with M antennas that serve K single-antenna PUs, where $M \geq K$. The cognitive network, on the other hand, consists of a cognitive BS equipped with N antennas and a CU with a single-antenna. For the primary network, the channel vector from the PBS to the k th PU is denoted by $\mathbf{h}_k = [h_{k,1}, \dots, h_{k,M}] \in \mathcal{C}^{1 \times M}$, where $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \gamma_k \mathbf{I}_M)$, $h_{k,i}$ denotes the small-scale fading coefficient from the i th PBS antenna to the k th PU and γ_k denotes the path loss to the k th PU.

The channel vector of the interfering link from the cognitive BS to the k th PU is denoted by $\mathbf{g}_k = [g_{k,1}, \dots, g_{k,N}] \in \mathcal{C}^{1 \times N}$, where $\mathbf{g}_k \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{I}_N)$, $g_{k,j}$ denotes the small-scale fading coefficient from the j th cognitive BS antenna to the k th PU and β_k denotes the path loss between cognitive BS and the k th PU. For the cognitive network, the channel vector from the cognitive BS to the CU is $\mathbf{h}_s = [h_1, \dots, h_N] \in \mathcal{C}^{1 \times N}$ and

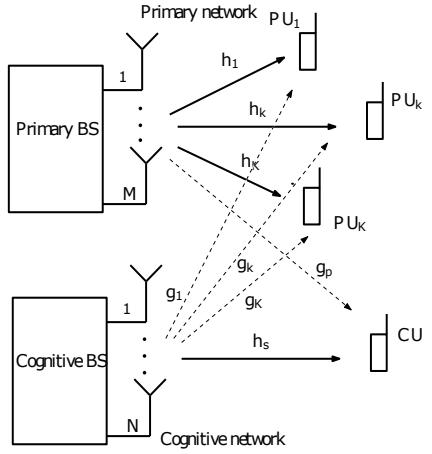


Fig. 1. Underlay cognitive massive MIMO network

$\mathbf{h}_s \sim \mathcal{CN}(\mathbf{0}, \alpha \mathbf{I}_N)$ where α captures the path loss. Furthermore, we denote the channel vector of the interfering link from the PBS to the CU by \mathbf{g}_p , where $\mathbf{g}_p = [g_1, \dots, g_M] \in \mathcal{C}^{1 \times M}$ and $\mathbf{g}_p \sim \mathcal{CN}(\mathbf{0}, \lambda \mathbf{I}_M)$. We assume that the channel vectors are mutually independent of each other, which is justified since the terminals are spatially separated.

A. Data Transmission

When the cognitive BS employs maximum ratio transmission to transmit a symbol x with power P , then the signal y_c received at the CU is given by

$$y_c = \sqrt{P} \mathbf{h}_s \frac{\mathbf{h}_s^\dagger}{\|\mathbf{h}_s^\dagger\|} x + \sqrt{E_p} \mathbf{g}_p \mathbf{W} \mathbf{u}^T + n, \quad (1)$$

where the data symbol x is chosen such that $\mathbb{E}[x] = 0$ and $\mathbb{E}[|x|^2] = 1$. Furthermore, E_p denotes the average PBS transmit power, n denotes additive white gaussian noise (AWGN) at CU and has zero mean and variance σ_n^2 . The interfering symbol vector transmitted from the PBS is given by $\mathbf{u} = [u_1, \dots, u_K]$. It is chosen such that $\mathbb{E}[\mathbf{u}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{u}^\dagger \mathbf{u}] = \mathbf{I}_K$. The maximum ratio precoding matrix at the PBS is denoted by $\mathbf{W} \in \mathcal{C}^{M \times K}$ and is given by $\mathbf{W} = \sqrt{\varepsilon} \mathbf{H}$. The k th column of \mathbf{H} corresponds to \mathbf{h}_k^T and the normalization constant ε is chosen such that the power constraint $\mathbb{E}[\text{tr}(\mathbf{W}^\dagger \mathbf{W})] = 1$ is satisfied at the PBS. Thus, $\varepsilon = \frac{1}{\mathbb{E}[\text{tr}(\mathbf{H}^\dagger \mathbf{H})]}$. Due to transmissions by the cognitive BS, the power I_{s_k} of the interference caused to the k th PU is $I_{s_k} = P \left| \frac{\mathbf{g}_k \mathbf{h}_s^\dagger}{\|\mathbf{h}_s^\dagger\|} \right|^2$, $1 \leq k \leq K$.

B. Interference Violation Probability Constraint

In this work, we consider an underlay cognitive massive MIMO network in which the cognitive BS is subject to an interference violation probability constraint. This constraint mandates that the interference power due to transmissions by the cognitive BS does not exceed a threshold I_p at any of the K PUs more than P_o fraction of the time. Mathematically, this can be written as

$$\Pr(I_{s_1} \leq I_p, I_{s_2} \leq I_p, \dots, I_{s_K} \leq I_p) \geq (1 - P_o), \quad (2)$$

where P_o is referred to as the interference violation probability. Please note that this constraint is a generalization of the peak interference constraint for which $P_o = 0$. Moreover, this practically motivated constraint is widely used to design primary exclusive zones to protect the PUs in CR networks [11] and requires knowledge of the channel states from the cognitive BS to the K PUs.

1) *Perfect CSI*: If the cognitive BS has perfect CSI of its channels to the K PUs and adapts its power as follows:

$$P = \frac{I_p}{\max_k \left\{ \left| \mathbf{g}_k \frac{\mathbf{h}_s^\dagger}{\|\mathbf{h}_s^\dagger\|} \right|^2 \right\}}, \quad (3)$$

it can be easily shown that P_o equals zero and $\Pr(I_{s_1} \leq I_p, I_{s_2} \leq I_p, \dots, I_{s_K} \leq I_p) = 1$.

2) *Imperfect CSI*: In practical wireless systems, the channels are estimated and thus known imperfectly. With imperfect CSI, P_o will not be zero with the power adaptation policy in (3), where the true channel vectors are replaced by their estimates. In this work, we use the imperfect CSI model based on Gauss-Markov uncertainty [12].

Based on this model, the true channel vector \mathbf{g}_k between the cognitive BS and the k th PU can be written as

$$\mathbf{g}_k = s_p \hat{\mathbf{g}}_k + \left(\sqrt{1 - s_p^2} \right) \mathbf{e}_k, \quad (4)$$

where $\hat{\mathbf{g}}_k \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{I}_N)$ is the estimate of the channel vector, $\mathbf{e}_k \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{I}_N)$ denotes the Gaussian noise which is uncorrelated to $\hat{\mathbf{g}}_k$ and s_p captures channel estimation error. Please note that $s_p = 1$ corresponds to perfect CSI and $0 < s_p < 1$ corresponds to imperfect CSI.

Similarly, the true channel \mathbf{h}_s between the cognitive BS and the CU can be written as

$$\mathbf{h}_s = s_s \hat{\mathbf{h}}_s + \left(\sqrt{1 - s_s^2} \right) \mathbf{e}_s, \quad (5)$$

where $\hat{\mathbf{h}}_s \sim \mathcal{CN}(\mathbf{0}, \alpha \mathbf{I}_N)$ is the estimate of the channel vector, $\mathbf{e}_s \sim \mathcal{CN}(\mathbf{0}, \alpha \mathbf{I}_N)$ denotes the Gaussian noise term which is uncorrelated to $\hat{\mathbf{h}}_s$ and s_s captures the estimation error.

III. PROPOSED POWER POLICY AND INTERFERENCE VIOLATION PROBABILITY WITH IMPERFECT CSI

As mentioned above, with imperfect CSI, P_o will not be zero. However, P_o can be constrained to meet a certain target. We propose a back-off factor based power adaptation policy that is a function of channel estimates from the cognitive BS to the K PUs and to the CU. Under this policy, the cognitive BS transmits with power

$$\hat{P} = \frac{I_p}{\eta \max_k \left\{ \left| \hat{\mathbf{g}}_k \frac{\hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|} \right|^2 \right\}}, \quad (6)$$

where the back-off factor η is chosen such that the interference violation probability constraint under imperfect CSI is met with equality for a target violation probability P_o . In other

words, given P_o , K , and s_p , the back-off factor η is chosen such that

$$\Pr(\hat{I}_{s_1} \leq I_p, \hat{I}_{s_2} \leq I_p, \dots, \hat{I}_{s_K} \leq I_p) = 1 - P_o, \quad (7)$$

$$\text{where } \hat{I}_{s_k} = \hat{P} \left| \mathbf{g}_k \frac{\hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|} \right|^2.$$

We state below a lower bound on the complement of the interference violation probability.¹

Theorem 1: A lower bound (P_L) on the complement of interference violation probability with imperfect CSI is

$$(1 - P_o) \geq P_L = 1 - K \eta B(\eta, K) + K \sum_{k=1}^L \left(1 - \exp\left(\frac{-x_k}{\eta}\right) \right)^{K-1} \times Q_1 \left(\sqrt{\frac{2s_p^2 x_k}{1 - s_p^2}}, \sqrt{\frac{2x_k}{\eta(1 - s_p^2)}} \right) w_{x_k}, \quad (8)$$

where $B(\cdot, \cdot)$ is the Beta function [13, (8.380.1)], $Q_1(\cdot, \cdot)$ is the first order Marcum-Q function [14, (4.34)], $\{x_k\}$ are the integration points, $\{w_{x_k}\}$ are the corresponding weights obtained via Gauss-Laguerre integration and L is the total number of integration points.

Proof: The proof is given in Appendix A. ■

Remark: It can be observed that P_o depends on η , K , s_p , and is independent of s_s and the number of cognitive BS antennas N . Based on Theorem 1, the wireless system designer can choose a rough ballpark number for η in order to satisfy the target interference violation probability at the PUs.²

IV. ACHIEVABLE RATE ANALYSIS WITH IMPERFECT CSI

In this section, we derive a lower bound on the achievable rate of the CU. With imperfect CSI and with the proposed power adaptation policy, the SINR ($\Gamma(\eta, P_o, s_p, s_s, K)$) at the CU can be written as

$$\Gamma(\eta, P_o, s_p, s_s, K) = \frac{\frac{I_p}{\eta \max_k \left\{ \left| \hat{\mathbf{g}}_k \frac{\hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|} \right|^2 \right\}} s_s^2 \|\hat{\mathbf{h}}_s^\dagger\|^2}{\varepsilon E_P Z + \alpha(1 - s_s^2) \frac{I_p}{\eta \max_k \left\{ \left| \hat{\mathbf{g}}_k \frac{\hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|} \right|^2 \right\}} + \sigma_n^2} \quad (9)$$

where $Z = \|\mathbf{g}_p \mathbf{H}\|^2$. Note that the SINR at the CU is a function of I_p , s_s and the back-off factor η , which in turn depends on s_p , K , and the target P_o . The achievable rate of the CU is stated below and obtained in a simplified form for $\gamma_k = \gamma$ for all k .

Theorem 2: The downlink achievable rate R_s of the CU with imperfect CSI and a back-off factor based power adaptation policy while ensuring that the cognitive transmissions violate the interference violation probability constraint no more than P_o fraction of the time is given by

$$R_s = \mathbb{E}[\log_2(1 + \Gamma(\eta, P_o, s_p, s_s, K))], \quad (10)$$

$$\geq \log_2(1 + e^{\mathbb{E}[\ln \Gamma(\eta, P_o, s_p, s_s, K)]}), \quad (11)$$

¹To ensure mathematical tractability and to gain insights, we focus on the case where the channel vectors from the cognitive BS to the K PUs are statistically identical, i.e. $\beta_k = \beta$ for all k .

²In the analysis that ensues, in order to obtain insights, we assume that the PBS knows its channels perfectly to the K PUs and the CU.

where $\mathbb{E}[\ln \Gamma(\eta, P_o, s_p, s_s, K)] = T_1 - T_2$,

$$T_1 = K(-1)^{(K-1)} \sum_{k=0}^{K-1} (-1)^k \binom{K-1}{k} \times \frac{1}{(K-k)} \left(\sum_{m=1}^{N-1} \frac{1}{m} - \ln \left(\frac{\eta\beta}{\alpha I_p s_s^2 (K-k)} \right) \right), \quad (12)$$

$$T_2 = \frac{2K(\lambda\gamma)^{\frac{-(M+K)}{2}}}{\eta\beta(M-1)!(K-1)!} \int_0^\infty \int_0^\infty \ln \left(\varepsilon E_P z + \frac{\alpha(1-s_s^2)I_p}{q_1} + \sigma_n^2 \right) \times \left(1 - e^{-\frac{q_1}{\eta\beta}} \right)^{K-1} e^{\frac{-q_1}{\eta\beta}} z^{\frac{K+M-2}{2}} K_\nu \left(2\sqrt{\frac{z}{\lambda\gamma}} \right) dz dq_1, \quad (13)$$

where $K_\nu(\cdot)$ is modified Bessel function of second kind with order $\nu = K - M$ [13, 8.407.1].

Proof: The proof is given in Appendix B. \blacksquare

We invoke Jensen's inequality in (10) to obtain a lower bound in (11) on the achievable rate. We evaluate T_2 numerically as it cannot be simplified any further.

Corollary 1: If the cognitive BS has perfect CSI of its channels to the CU and $\sigma_n^2 = 1$, T_2 in (13) reduces to

$$T'_2 = \frac{1}{(M-1)!(K-1)!} G_{4,2}^{1,4} \left[\varepsilon E_P \lambda \gamma \middle| \begin{matrix} (1-K), (1-M), 1, 1 \\ 1, 0 \end{matrix} \right], \quad (14)$$

where $G_{p,q}^{m,n} \left[x \middle| \begin{matrix} a_1, \dots, a_p \\ b_1, \dots, b_q \end{matrix} \right]$ is Meijer G function [13, 9.301].

Proof: Using $\ln(\varepsilon E_P z + 1) = G_{2,2}^{1,2} \left[\varepsilon E_P z \middle| \begin{matrix} 1, 1 \\ 1, 0 \end{matrix} \right]$ and the identity in [13, (7.821.3)], we obtain (14). \blacksquare

V. NUMERICAL RESULTS AND DISCUSSION

In this section, we present numerical results to understand the interplay among various system parameters such as η , s_p , s_s , and K on the complement of the interference violation probability. We then investigate the use of massive MIMO in reducing P_o at the PUs without degrading the rate of the CU and to illustrate the effects of K and P_o on the rate of the CU. For illustration, we take $\beta = \alpha = \gamma = \lambda = 1$.

Figure 2 plots the complement of interference violation probability $(1 - P_o)$ vs. the back-off factor η for $K = 10$, $N = 100$, and for different values of s_p . We obtain exact curves through Monte Carlo simulation of (7). We also plot the lower bound obtained through (8) (for $L = 6$). Note that it is fairly close to the exact curve for $P_o \leq 20\%$, which is essentially the operational regime. As η increases, P_o reduces since the cognitive BS transmits at a reduced power. Furthermore, as s_p increases, the channel estimation accuracy increases. Hence, the cognitive BS violates the interference constraint less often. In other words, the larger the estimation error, the greater is the amount of back-off needed in power to maintain the violation probability at the same level. Note that η is unaffected by s_s for a given P_o , since the power with

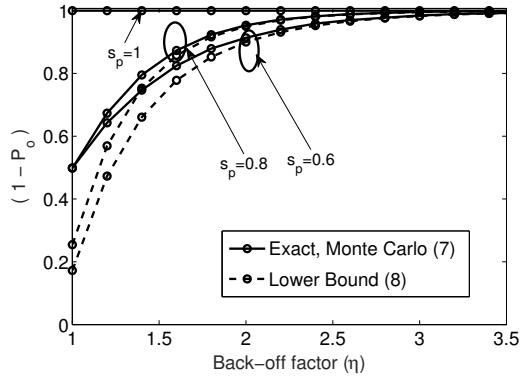


Fig. 2. $(1 - P_o)$ vs. η for different s_p , $K = 10$, and $N = 100$

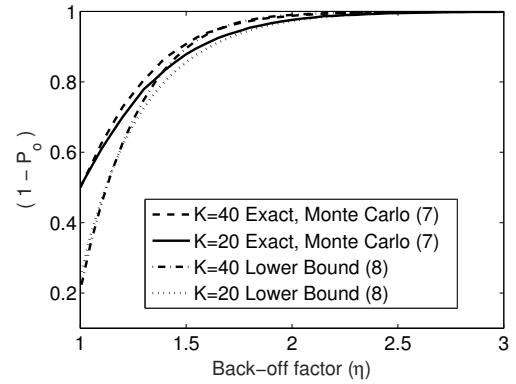


Fig. 3. $(1 - P_o)$ vs. η for different K , $s_p = 0.8$, $N = 100$

which the cognitive BS transmits depends only on the strength of the channels from the cognitive BS to the PUs. Note that $s_p = 1$ corresponds to perfect CSI and $P_o = 0$ for this case.

Figure 3 plots $(1 - P_o)$ vs. η for $s_p = 0.8$, $N = 100$, and for different values of K . For a fixed value of η , the interference violation probability P_o is equal to or lower with $K = 40$ when compared to $K = 20$. This is because the maximum of a larger set ($K = 40$) is always greater than or equal to the maximum of a smaller set ($K = 20$). Hence, the cognitive BS transmits at a lower power with $K = 40$ than with $K = 20$.

Figure 4 plots R_s (both exact (10) and the lower bound ((11)-(14))) vs. N for different $(\eta, 1 - P_o)$ sampled from Figure 2, $s_s = 1$ and 0.4 , $s_p = 0.8$, $K = 10$, $M = 100$, $I_p = 10$ dB, and $E_p = 10$ dB. For a specific s_p and s_s , P_o can be reduced while maintaining R_s at a constant level by deploying more number of antennas at the cognitive BS. For example, by increasing N from 60 to 77, we can reduce P_o from 20% to 8% while maintaining a constant achievable rate of 4.018 bits/s/Hz for the CU with $s_s = 1$. The reduction in power due to a larger back-off is compensated by deploying more antennas at the cognitive BS. Furthermore, as s_s decreases, the quality of the channel estimates degrade and the rate of the CU drops. As mentioned earlier in Figure 2, for a given P_o , η is unaffected by s_s .

Figure 5 plots R_s (both exact (10) and the lower

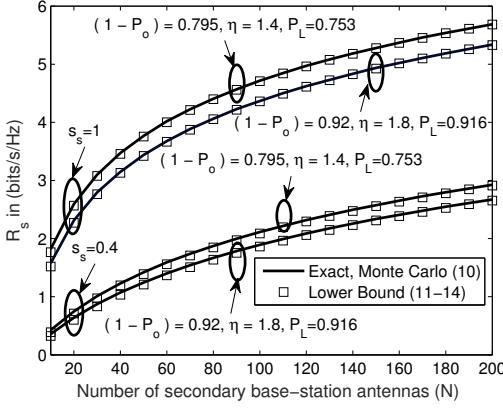


Fig. 4. R_s vs. N for different $(\eta, 1 - P_o)$, $s_p = 0.8$, $\sigma_n^2 = 1$, $K = 10$, $M = 100$, $I_p = 10$ dB, $E_p = 10$ dB

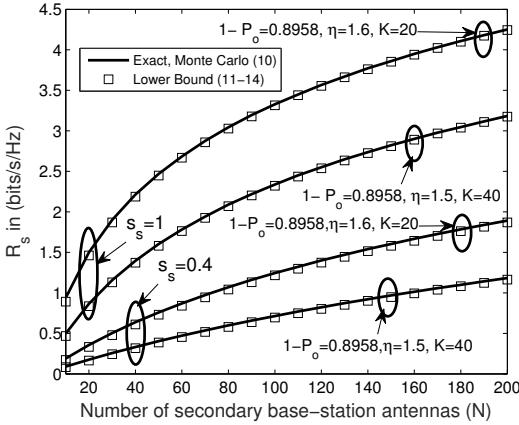


Fig. 5. R_s vs. N for different (K, η) , $s_p = 0.8$, $\sigma_n^2 = 1$, $(1 - P_o) = 0.8958$, $M = 100$, $I_p = 10$ dB, $E_p = 10$ dB

bound ((11)-(14)) vs. N for different (K, η) sampled from Figure 3 for $(1 - P_o) = 0.8958$, $s_s = 1$ and 0.4 , $s_p = 0.8$, $M = 100$, $I_p = 10$ dB, and $E_p = 10$ dB. We observe that R_s decreases as we increase K . This is because the cognitive BS transmits at a lower power as K increases. As before, R_s also decreases as s_s decreases, since the channel estimation error increases with reduction in s_s . Another important observation is that even if the number of PUs increases, $(1 - P_o)$ and the rate of the CU can be maintained at the same level by increasing the number of cognitive BS antennas. For example, for $s_s = 1$ and $P_o = 10\%$, we can accommodate 20 more PUs in the network while maintaining rate of CU at 1.9 bits/s/Hz by increasing N from 30 to 70.

VI. CONCLUSIONS

We studied an underlay cognitive massive MIMO network in which the cognitive BS has an imperfect estimate of its channels to the K PUs and the CU and is subject to an interference violation probability constraint. Given this set up, we proposed a simple back-off factor based power adaptation policy for the cognitive BS such that it exceeds an interference threshold I_p no more than P_o fraction of the time at any of

the K PUs. The lower bound on $(1 - P_o)$ gives the wireless system designer a ballpark number for η that must be chosen to satisfy the constraint. We then derived a lower bound on the achievable rate of the CU. We observed and quantified that by exploiting the huge array gain and the higher spatial resolution offered by a massive MIMO cognitive BS, P_o can be reduced without incurring a degradation in the rate of the CU. Also, more number of PUs can be served while keeping P_o at the PUs and rate of the CU fixed at a constant level by deploying a higher N . Analysis with multiple CUs is an interesting avenue for future work.

APPENDIX

A. Brief Proof of Theorem I

Let $u_k = \frac{\mathbf{g}_k \hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|}$ and $\hat{u}_k = \frac{\hat{\mathbf{g}}_k \hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|}$ for $1 \leq k \leq K$. It can be shown that u_k and \hat{u}_k conditioned on $\hat{\mathbf{h}}_s^\dagger$ are $\mathcal{CN}(0, \beta_k)$ rvs and are independent of $\hat{\mathbf{h}}_s$. Furthermore, let $X_k = |u_k|^2$ and $\hat{X}_k = |\hat{u}_k|^2$ where X_k and \hat{X}_k are correlated exponential rvs. Thus, (7) can be written as

$$(1 - P_o) = \Pr(X_1 \leq \eta \max_k \{\hat{X}_k\}, \dots, X_K \leq \eta \max_k \{\hat{X}_k\}) \quad (15)$$

Using Bonferroni's inequality [15] and for $\beta_k = \beta$ for all k , a lower bound on (15) is

$$\begin{aligned} (1 - P_o) &\geq K \Pr(X_1 \leq \eta \max_k \{\hat{X}_k\}) - (K - 1) \\ &= 1 - K \left[\Pr\left(\hat{X}_1 \leq \frac{X_1}{\eta}, \dots, \hat{X}_K \leq \frac{X_1}{\eta}\right) \right]. \end{aligned}$$

This can be rewritten as

$$(1 - P_o) \geq 1 - K \int_{x_1=0}^{\infty} \int_{\hat{x}_1=0}^{\frac{x_1}{\eta}} \left(1 - e^{-\frac{x_1}{\eta\beta}}\right)^{K-1} f_{X_1, \hat{X}_1}(x_1, \hat{x}_1) d\hat{x}_1 dx_1 \quad (16)$$

where $f_{X_1, \hat{X}_1}(x_1, \hat{x}_1)$ is the bivariate exponential pdf which can be obtained through a simple rv transformation and using [14, (6.2)]. Upon substituting the joint pdf of X_1 and \hat{X}_1 and rearranging, we get $(1 - P_o) \geq 1 - I_1 + I_2$ where

$$\begin{aligned} I_1 &= \frac{K}{\beta^2(1 - s_p^2)} \int_{x_1=0}^{\infty} \left(1 - e^{-\frac{x_1}{\eta\beta}}\right)^{K-1} \exp\left(\frac{-x_1}{\beta(1 - s_p^2)}\right) \\ &\times \int_{\hat{x}_1=0}^{\infty} \exp\left(\frac{-\hat{x}_1}{\beta(1 - s_p^2)}\right) I_0\left(\frac{2s_p\sqrt{x_1\hat{x}_1}}{\beta(1 - s_p^2)}\right) d\hat{x}_1 dx_1 \quad (17) \end{aligned}$$

$$\begin{aligned} I_2 &= \frac{K}{\beta^2(1 - s_p^2)} \int_{x_1=0}^{\infty} \left(1 - e^{-\frac{x_1}{\eta\beta}}\right)^{K-1} \exp\left(\frac{-x_1}{\beta(1 - s_p^2)}\right) \\ &\times \int_{\hat{x}_1=\frac{x_1}{\eta}}^{\infty} \exp\left(\frac{-\hat{x}_1}{\beta(1 - s_p^2)}\right) I_0\left(\frac{2s_p\sqrt{x_1\hat{x}_1}}{\beta(1 - s_p^2)}\right) d\hat{x}_1 dx_1 \quad (18) \end{aligned}$$

I_1 in (17) can further be simplified using [13, (6.614.3)] as

$$\begin{aligned} I_1 &= \frac{K\sqrt{1 - s_p^2}}{\sqrt{\beta}s_p} \int_{x_1=0}^{\infty} x_1^{-\frac{1}{2}} \left(1 - \exp\left(-\frac{x_1}{\eta\beta}\right)\right)^{K-1} \\ &\times \exp\left(\frac{x_1(s_p^2 - 2)}{2\beta(1 - s_p^2)}\right) M_{-\frac{1}{2}, 0}\left(\frac{s_p^2 x_1}{\beta(1 - s_p^2)}\right) dx_1 \quad (19) \end{aligned}$$

where $M_{\lambda,\mu}(z)$ is Whittaker function which can be written in terms of Confluent hypergeometric function $\Phi(a, b; x)$ [13, (9.220.2)] as

$$I_1 = \frac{K}{\beta} \int_0^\infty \left(1 - e^{-\frac{x_1}{\eta\beta}}\right)^{K-1} e^{\frac{-x_1}{\beta(1-s_p^2)}} \Phi\left(1, 1; \frac{s_p^2 x_1}{\beta(1-s_p^2)}\right) dx_1 \quad (20)$$

Applying [13, (3.312.1) and (9.215.1)], $I_1 = K\eta B(\eta, K)$, where $B(\cdot, \cdot)$ is the Beta function [13, (8.380.1)]. To simplify I_2 in (18), we solve the inner integral by substituting $\frac{\hat{x}_1}{\beta(1-s_p^2)} = \frac{y_2}{2}$ and using [14, (4.34)] to obtain

$$I_2 = \frac{K}{\beta} \int_0^\infty e^{\frac{-x_1}{\beta}} \left(1 - e^{-\frac{x_1}{\eta\beta}}\right)^{K-1} Q_1\left(\sqrt{\frac{2s_p^2 x_1}{\beta(1-s_p^2)}}, \sqrt{\frac{2x_1}{\eta\beta(1-s_p^2)}}\right) dx_1 \quad (21)$$

where, $Q_1(\cdot, \cdot)$ is first order Marcum-Q function. We then obtain closed-form for I_2 using Gauss-Laguerre integration.

B. Brief Proof of Theorem 2

We are interested in computing

$$\begin{aligned} & \mathbb{E}[\ln \Gamma(\eta, P_o, s_p, s_s, K)] \\ &= \mathbb{E}[\ln X] - \mathbb{E}\left[\ln\left(\varepsilon E_P Z + \alpha(1 - s_s^2) \frac{I_p}{Q_1} + \sigma_n^2\right)\right] \end{aligned} \quad (22)$$

where $X = \frac{I_p}{Q_1} s_s^2 Q_2 = I_p s_s^2 U$ and $U = \frac{Q_2}{Q_1}$. The rv $Q_1 = \eta \max_k |Y_k|^2$ where $Y_k = \hat{\mathbf{g}}_k \frac{\hat{\mathbf{h}}_s^\dagger}{\|\hat{\mathbf{h}}_s^\dagger\|}$, and $Q_2 = \|\hat{\mathbf{h}}_s^\dagger\|^2 = |\hat{h}_1^\dagger|^2 + |\hat{h}_2^\dagger|^2 + \dots + |\hat{h}_N^\dagger|^2$. We need to compute the pdf of U which is the ratio of Q_2 and Q_1 .

It is easy to show that the Y_k conditioned on $\hat{\mathbf{h}}_s^\dagger$ is independent of $\hat{\mathbf{h}}_s^\dagger$ and is a $\mathcal{CN}(0, \beta)$ rv. Therefore, Q_1 being a function of Y_k is independent of $\hat{\mathbf{h}}_s^\dagger$. Hence, it is independent of Q_2 . Furthermore, Q_1 is η times the maximum of independent and identically distributed (i.i.d) exponential rvs, its CDF is $F_{Q_1}(q_1) = \left(1 - \exp\left(-\frac{q_1}{\eta\beta}\right)\right)^K$. Therefore, the pdf of Q_1 is

$$f_{Q_1}(q_1) = \frac{K}{\eta\beta} \left(1 - \exp\left(-\frac{q_1}{\eta\beta}\right)\right)^{K-1} \exp\left(\frac{-q_1}{\eta\beta}\right), \quad q_1 > 0 \quad (23)$$

We next find the pdf of Q_2 . Since, $\frac{2Q_2}{\alpha}$ is the sum of squares of $(2N)$ i.i.d Gaussian rvs with zero mean and unit variance, it is chi-square distributed rv with $2N$ degrees of freedom and its pdf can be written through a simple rv transformation as

$$f_{Q_2}(q_2) = \frac{1}{\alpha^N (N-1)!} q_2^{(N-1)} e^{-\frac{q_2}{\alpha}}, \quad q_2 > 0 \quad (24)$$

Using the pdfs of Q_1 and Q_2 from (23) and (24) respectively, and using [13, (3.432.1)], the pdf of U is obtained as

$$\begin{aligned} f_U(u) &= \eta^N N K (-1)^{(K-1)} (\beta/\alpha)^N u^{(N-1)} \sum_{k=0}^{K-1} (-1)^k \\ &\times \binom{K-1}{k} \frac{1}{(K + \frac{\eta\beta u}{\alpha} - k)^{(N+1)}}, \quad u > 0 \end{aligned} \quad (25)$$

Thus,

$$\begin{aligned} \mathbb{E}[\ln(I_p s_s^2 U)] &= \eta^N N K (-1)^{(K-1)} \left(\frac{\beta}{\alpha}\right)^N \sum_{k=0}^{K-1} (-1)^k \binom{K-1}{k} \\ &\times \int_0^\infty u^{(N-1)} \ln(I_p s_s^2 u) \frac{1}{(K + \frac{\eta\beta u}{\alpha} - k)^{(N+1)}} du, \end{aligned} \quad (26)$$

Substituting $\ln(I_p s_s^2 u) = t$, $t = -x$ and using [13, (3.458.2)], (26) simplifies to T_1 in (12).

To compute $\mathbb{E}\left[\ln\left(\varepsilon E_P Z + \alpha(1 - s_s^2) \frac{I_p}{Q_1} + \sigma_n^2\right)\right]$, the pdf of $Z = \|\mathbf{g}_p \mathbf{H}\|^2$ is simplified using the result in [6] and [13, (3.471.9)] as

$$f_Z(z) = \frac{2z^{\frac{K+M-2}{2}} K_\nu\left(2\sqrt{z/(\lambda\gamma)}\right)}{(\lambda\gamma)^{(M+K)/2} (M-1)!(K-1)!}, \quad z > 0 \quad (27)$$

where $\nu = (K-M)$. Substituting the pdfs of Q_1 and Z from (23) and (27) respectively, the second term in (22) can be simplified to obtain T_2 as stated in Theorem 2.

REFERENCES

- [1] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894–914, May. 2009.
- [2] S. Kashyap and N. B. Mehta, "SEP-optimal transmit power policy for peak power and interference outage probability constrained underlay cognitive radios," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, pp. 6371–6381, Dec. 2013.
- [3] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [4] H. Xie, B. Wang, F. Gao, and S. Jin, "A full-space spectrum-sharing strategy for massive MIMO cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2537–2549, Oct. 2016.
- [5] G. Geraci, A. G.-Rodriguez, D. L.-Perez, A. Bonfante, G. Giordano, and H. Claussen, "Operating massive MIMO in unlicensed bands for enhanced coexistence and spatial reuse," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1282–1293, Jun. 2017.
- [6] L. Wang, H. Q. Ngo, M. Elkashlan, T. Q. Duong, and K.-K. Wong, "Massive MIMO in spectrum sharing networks: Achievable rate and power efficiency," *IEEE Systems Journal*, vol. 11, no. 1, pp. 20–31, Mar. 2017.
- [7] M. Cui, B. J. Hu, X. Li, H. Chen, S. Hu, and Y. Wang, "Energy-efficient power control algorithms in massive MIMO cognitive radio networks," *IEEE Access*, vol. 5, pp. 1164–1177, Jan. 2017.
- [8] S. Chaudhari and D. Cabric, "Downlink transceiver beamforming and admission control for massive MIMO cognitive radio networks," *Proc. Asilomar Conf. on Signals, Syst., and Comput.*, pp. 1257–1261, Nov. 2015.
- [9] W. Hao, O. Muta, H. Gacanin, and H. Furukawa, "Power allocation for massive MIMO cognitive radio networks with pilot sharing under SINR requirements of primary users," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1174–1186, Feb. 2018.
- [10] M. Cui, B.-J. Hu, J. Tang, and Y. Wang, "Energy-efficient joint power allocation in uplink massive MIMO cognitive radio networks with imperfect CSI," *IEEE Access*, vol. 5, pp. 27611–27621, Nov. 2017.
- [11] M. Vu, N. Devroye, and V. Tarokh, "On the primary exclusive region of cognitive networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3380–3385, Jul. 2009.
- [12] B. Nosrat-Makouei, J. G. Andrews, and R. W. Heath, "MIMO interference alignment over correlated channels with imperfect CSI," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2783–2794, June 2011.
- [13] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 4th ed. Academic Press, 1980.
- [14] M. Simon and M.-S. Alouini, *Digital Communication over Fading Channels*, 2nd ed. Wiley-Interscience, 2005.
- [15] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2nd ed. Athena Scientific, 2008.

Improved Tail Bounds for Missing Mass and Confidence Intervals for Good-Turing Estimator

Prafulla Chandra, Aditya Pradeep and Andrew Thangaraj

Department of Electrical Engineering
Indian Institute of Technology, Madras
Chennai, India 600036
{ee16d402, ee14b068, andrew}@ee.iitm.ac.in

Abstract—The missing mass of a sequence is defined as the total probability of the elements that have not appeared or occurred in the sequence. The popular Good-Turing estimator for missing mass has been used extensively in language modeling and ecological studies. Exponential tail bounds have been known for missing mass, and improving them results in better confidence in estimation. In this work, we first show that missing mass is sub-Gamma on the right tail with the best-possible variance parameter under the Poisson and multinomial sampling models. This results in a right tail bound that beats the previously best known tail bound for deviation from mean up to about 0.2785. Further, we show that the sub-Gaussian approach cannot result in any improvement in the right tail bound for Poisson sampling. We derive confidence intervals for the Good-Turing estimator with better confidence levels and narrower width when compared to existing ones. Our results are worst case over all distributions.

I. INTRODUCTION

A. Definitions and Notation

Let P be an arbitrary discrete distribution on an alphabet \mathcal{X} , and let N be a random variable taking non-negative integer values. Consider random samples $X^N = (X_1, X_2, \dots, X_N)$ with $X_i \sim P$ iid and N independent of X_i . In the multinomial sampling model, we have $N = n$ with probability 1 for a positive integer n . In the Poisson sampling model, we have $N \sim \text{Poisson}(n)$. We will assume that n is known in either model, but no assumptions will be made on P and \mathcal{X} . For $x \in \mathcal{X}$, let $p_x \triangleq P(x)$. Indicator random variables and expectations are denoted $I(\cdot)$ and $E[\cdot]$, respectively.

For $x \in \mathcal{X}$, let $F_x(X^N) = \sum_{i=1}^N I(X_i = x)$ denote the number of occurrences of x in X^N . For $k \geq 0$, let $\phi_k(X^N) = \sum_{x \in \mathcal{X}} I(F_x(X^N) = k)$ denote the number of letters that have occurred k times in X^N . For $k \geq 0$, the k -th combined probability mass, denoted $M_k(X^N, P)$, is defined as

$$M_k(X^N, P) \triangleq \sum_{x \in \mathcal{X}} p_x I(F_x(X^N) = k). \quad (1)$$

$M_0(X^N, P)$ is called the missing mass. We will drop the arguments X^N and P whenever possible. We will use normal notation F_x , ϕ_k and M_k when referring exclusively to multinomial sampling and bold notation \mathbf{F}_x , $\boldsymbol{\phi}_x$ and \mathbf{M}_k for Poisson sampling.

In this article, the Good-Turing estimator for the k -th combined probability mass [1] [2] [3] is defined as follows:

$$G_k(X^n) \triangleq \frac{(k+1)\phi_{k+1}(X^n)}{n-k} \quad (\text{multinomial}) \quad (2)$$

$$\mathbf{G}_k(X^N) \triangleq \frac{(k+1)\phi_{k+1}(X^N)}{n} \quad (\text{Poisson}) \quad (3)$$

Missing mass and the Good-Turing estimator have been extensively used in multiple applications and are important in theory and practice [2], [4]–[10]. The combined probability mass $M_k(X^N, P)$ and the error of the Good-Turing estimator $G_k(X^N) - M_k(X^N, P)$ concentrate around their expected value, and this has been exploited to derive exponential left/right tail bounds and confidence intervals [3], [11]–[13]. In this work, we make specific improvements to a few such previously best-known results.

B. Sub-Gaussian, Sub-Poisson and Sub-Gamma

For a random variable Z , let

$$L_Z(\lambda) \triangleq \ln(E[e^{\lambda(Z-E[Z])}]), \quad (4)$$

$$T_Z(\epsilon) \triangleq -\ln \Pr(Z - E[Z] \geq \epsilon). \quad (5)$$

Table I provides upper bounds on $L_Z(\lambda)$ under which Z is sub-Gaussian, sub-Poisson or sub-Gamma on the right tail and the resulting lower bound on $T_Z(\epsilon)$ [13] [14].

TABLE I
SUB-GAUSSIAN, SUB-POISSON, SUB-GAMMA RANDOM VARIABLES:
 v - VARIANCE FACTOR, c - SCALE PARAMETER.

Property	UB on $L_Z(\lambda)$	LB on $T_Z(\epsilon)$
sub-Gaussian(v)	$\lambda^2 v/2$, $\lambda \geq 0$	$\epsilon^2/2v$
sub-Poisson(v)	$(e^\lambda - \lambda - 1)v$, all λ	
sub-Gamma(v, c)	$\frac{\lambda^2 v}{2(1-c\lambda)}$, $\lambda \in [0, 1/c]$	$\left(1 + \frac{ce}{v} - \sqrt{1 + \frac{2ce}{v}}\right) \frac{v}{c^2}$

A random variable Z is sub-Gaussian(v) or sub-Gamma(v, c) on the left tail if $-Z$ is, respectively, sub-Gaussian(v) or sub-Gamma(v, c) on the right tail. The following lemma collects together standard results on the three properties [13] [14].

Lemma 1. 1) If Z is sub-Poisson(v), then $-Z$ is sub-Gaussian(v) and Z is sub-Gamma($v, 1/3$) on the right tail.

- 2) If Z is sub-Gaussian(v), then Z is sub-Gamma(v, c) for any $c > 0$.
- 3) If Z is sub-Gaussian(v), sub-Poisson(v) or sub-Gamma(v, c), then $v \geq \text{Var}(Z)$.

C. Previous results

1) *Tail bounds:* The following right tail bound was proved in [11] under the multinomial sampling model:

$$\Pr(M_0 - E[M_0] \geq \epsilon) \leq e^{-n\epsilon^2}, \quad \epsilon \in [0, 1]. \quad (6)$$

In [12], under the multinomial sampling model, the following left tail bound was proved:

$$\Pr(E[M_0] - M_0 \geq \epsilon) \leq e^{-1.92n\epsilon^2}, \quad \epsilon \in [0, 1]. \quad (7)$$

The above two tail bounds of [11] and [12] are the current best and can be shown to hold for the Poisson sampling model as well.

2) *Confidence intervals:* In [3], the following was shown under the multinomial sampling model for $k \geq 0$:

With probability at least $1 - \delta$, it holds that

$$\begin{aligned} |G_k - M_k| &\leq \frac{k+2}{n-k} + \sqrt{\frac{2 \ln(\frac{3}{\delta})}{n}} \left[\frac{k+1}{1-\frac{k}{n}} + k \right] \\ &\quad + \sqrt{\frac{2 \ln(\frac{3}{\delta})}{n}} \left[\sqrt{2k \ln\left(\frac{3n}{\delta}\right)} + 2 \ln\left(\frac{3n}{\delta}\right) \right]. \end{aligned} \quad (8)$$

For fixed k and large n , the RHS above is $O\left(\frac{\ln n}{\sqrt{n}}\right)$. In [13], under the Poisson sampling model, the following was shown:

With probability at least $1 - 4\delta$, it holds that

$$G_0 \leq M_0 + \frac{\sqrt{2(\phi_1 + 2\phi_2) \ln(1/\delta)}}{n} + \frac{4 \ln(1/\delta)}{n}, \quad (9)$$

$$G_0 \geq M_0 - \frac{\sqrt{6(\phi_1 + \phi_2 + \dots) \ln(1/\delta)}}{n} - \frac{5 \ln(1/\delta)}{n}. \quad (10)$$

The interval is a random quantity. Using the bound $\phi_k \leq N/k$ and concentration of N around its mean n , order estimates for the intervals of (9) and (10) are $\sqrt{1/n}$ and $\sqrt{(\ln n)/n}$, which will hold with high probability for large n .

D. Our Results

1) *Sub-Poisson, sub-Gamma, sub-Gaussian properties:* Let γ and γ' be defined as

$$\gamma = \max_{t>0} t e^{-t} (1 - e^{-t}) \approx 0.2603, \quad (11)$$

$$\gamma' = \max_{t>0} 2t^2 e^{-t} \approx 1.083. \quad (12)$$

Let the error of the Good-Turing estimator be denoted $\mathbf{E}_k = \mathbf{G}_k - \mathbf{M}_k$. The following lemma summarizes our results.

Lemma 2. 1) In the Poisson sampling model, \mathbf{M}_0 is sub-Gamma($\gamma/n, 1/n$).
2) In the multinomial model, \mathbf{M}_0 is sub-Gamma($\gamma/n + \gamma'/n^2, 1/n$).

- 3) In the Poisson sampling model, there exists P such that \mathbf{M}_0 is not sub-Gaussian(c/n) for any $c \leq 1/2$.
- 4) In the Poisson sampling model, for $k \leq n-1$, $-\mathbf{E}_k$ is sub-Gaussian($\text{Var}(\mathbf{E}_k)$).
- 5) In the Poisson sampling model, \mathbf{E}_k is sub-Gamma($\text{Var}(\mathbf{E}_k), \frac{k+1}{3n}$).

A proof of the above is presented in Section II. Using Lemma 2, it is easy to derive tail bounds and confidence intervals. Part 4 of Lemma 2 is a left tail result, while all other parts are for the right tail. Part 5 extends a similar result in [13] for $k=0$ to higher values of k but improves the scaling factor by a factor of 3.

2) Right tail bounds:

Theorem 3. Under the Poisson sampling model, the following bound holds:

$$\Pr(\mathbf{M}_0 - E[\mathbf{M}_0] \geq \epsilon) \leq e^{-n(\epsilon + \gamma - \sqrt{\gamma(2\epsilon + \gamma)})}. \quad (13)$$

Proof. Use Part 1 of Lemma 2 and the sub-Gamma tail bound in Table I. \square

Theorem 4. Under the multinomial sampling model, the following bound holds:

$$\Pr(M_0 - E[M_0] \geq \epsilon) \leq e^{-n(\epsilon + \gamma + \frac{\gamma'}{n} - \sqrt{(\gamma + \frac{\gamma'}{n})(2\epsilon + \gamma + \frac{\gamma'}{n})})}. \quad (14)$$

Proof. Use Part 2 of Lemma 2 and the sub-Gamma tail bound in Table I. \square

The tail bound in (13) is lower than the previously known bound in (6) if

$$\begin{aligned} \epsilon + \gamma - \sqrt{\gamma(2\epsilon + \gamma)} &> \epsilon^2, \text{ or} \\ \epsilon \in [0, 1 - \sqrt{2\gamma}], \quad 1 - \sqrt{2\gamma} &\approx 0.2785. \end{aligned} \quad (15)$$

By a similar computation, the multinomial sampling tail bound in (14) is lower if

$$\epsilon \in [0, 1 - \sqrt{2(\gamma + \gamma'/n)}]. \quad (16)$$

To illustrate the comparison, Fig. 1 shows a plot of the bounds in (6), (13) and (14) for $n = 20, 100, 1000$.

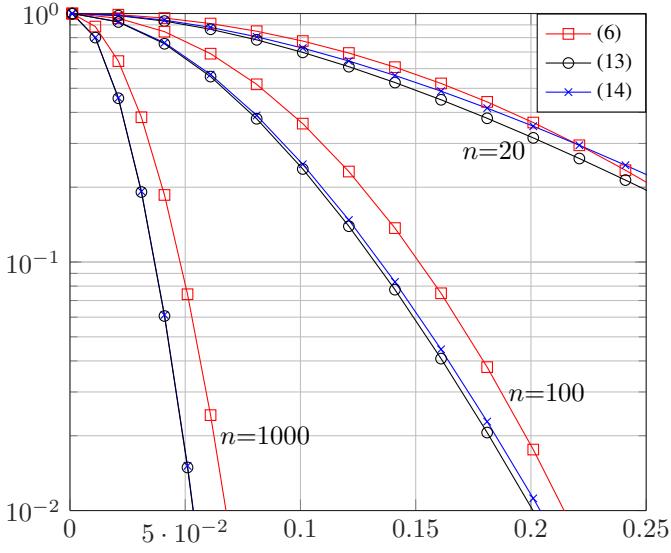


Fig. 1. Comparison of right tail bounds.

3) *Confidence intervals:* Using the fourth and fifth results in Lemma 2, we have the following.

Theorem 5. Under the Poisson sampling model, for $k \leq n-1$, with probability at least $1 - 3\delta$, it holds that

$$\mathbf{G}_k \leq \mathbf{M}_k + \frac{(k+1)\sqrt{2\left(\phi_{k+1} + \frac{k+2}{k+1}\phi_{k+2}\right)\ln(1/\delta)}}{n} + \frac{\left(\frac{k+1}{3} + 2\sqrt{(k+2)(k+1)}\right)\ln(1/\delta)}{n}, \quad (17)$$

$$\mathbf{G}_k \geq \mathbf{M}_k - \frac{(k+1)\sqrt{2\left(\phi_{k+1} + \frac{k+2}{k+1}\phi_{k+2}\right)\ln(1/\delta)}}{n} - \frac{2\sqrt{(k+2)(k+1)}\ln(1/\delta)}{n}. \quad (18)$$

Theorem 5 improves upon (9) and (10) in two ways - (1) the probability is improved from $1 - 4\delta$ to $1 - 3\delta$, (2) the interval estimate is reduced to $\sqrt{1/n}$ with high probability for large n . A proof for Theorem 5 is given in the next section.

II. PROOFS OF LEMMA 2 AND THEOREM 5

A proof for each part of Lemma 2 and a proof for Theorem 5 are given in this section. The following results are needed in the proofs and they are stated without proof.

Lemma 6. In the Poisson sampling model,

1) $\mathbf{F}_x \sim \text{Poisson}(np_x)$ and

$$E[\mathbf{M}_k] = \sum_{x \in \mathcal{X}} p_x e^{-np_x} \frac{(np_x)^k}{k!}, \quad (19)$$

$$\text{Var}(\mathbf{M}_0) \leq \frac{\gamma}{n}. \quad (20)$$

2) \mathbf{F}_x are independent and

$$L_{\mathbf{M}_0}(\lambda) = \sum_{x \in \mathcal{X}} L_{I(\mathbf{F}_x=0)}(\lambda p_x), \quad (21)$$

$$L_{\mathbf{E}_k}(\lambda) = \sum_{x \in \mathcal{X}} L_{\mathbf{E}_x}(\lambda), \quad (22)$$

where $\mathbf{E}_x = \frac{k+1}{n} I(\mathbf{F}_x = k+1) - p_x I(\mathbf{F}_x = k)$.

3) The mean (bias) and variance of the Good-Turing estimator error \mathbf{E}_k are as follows:

$$E[\mathbf{E}_k] = 0, \quad (23)$$

$$\text{Var}(\mathbf{E}_k) = \left(\frac{k+1}{n}\right)^2 E\left[\phi_{k+1} + \frac{k+2}{k+1}\phi_{k+2}\right]. \quad (24)$$

Lemma 7. In the multinomial sampling model,

1) $F_x \sim \text{Binomial}(n, p_x)$ and

$$E[M_k] = \sum_{x \in \mathcal{X}} \binom{n}{k} p_x^{k+1} (1-p_x)^{n-k}, \quad (25)$$

$$\text{Var}(M_0) \leq \frac{\gamma}{n} + \frac{\gamma'}{n}. \quad (26)$$

2) F_x are negatively associated [15] and

$$L_{M_0}(\lambda) \leq \sum_{x \in \mathcal{X}} L_{I(F_x=0)}(\lambda p_x). \quad (27)$$

A random variable X with $|X - E[X]| \leq 1$ satisfies the following inequality known as Bennett's inequality [14]:

$$L_X(\lambda) \leq \text{Var}(X)(e^\lambda - \lambda - 1), \quad \text{all } \lambda. \quad (28)$$

A. Sub-Gamma in Poisson sampling (Lemma 2, Part 1)

Using Bennett's inequality in (28) with $X = I(\mathbf{F}_x = 0)$ and setting λ as λp_x , we get

$$L_{I(F_x=0)}(\lambda p_x) \leq e^{-np_x} (1 - e^{-np_x})(e^{\lambda p_x} - \lambda p_x - 1). \quad (29)$$

In [13], the RHS above is further simplified using $1 - e^{-np_x} \leq 1$ and \mathbf{M}_0 was shown to be sub-Gamma on the right tail with variance factor $\frac{2\mathbb{E}[\phi_2]}{n^2}$ and scale parameter $\frac{1}{n}$. We proceed differently. Using the above inequality in (21), we get

$$\begin{aligned} L_{\mathbf{M}_0}(\lambda) &\leq \sum_{x \in \mathcal{X}} e^{-np_x} (1 - e^{-np_x})(e^{\lambda p_x} - \lambda p_x - 1) \\ &= \sum_{r=2}^{\infty} \frac{\lambda^r}{r! n^{r-1}} \left(\sum_{x \in \mathcal{X}} p_x e^{-np_x} (1 - e^{-np_x}) (np_x)^{r-1} \right) \\ &= \sum_{r=2}^{\infty} \frac{\lambda^r}{r! n^{r-1}} \sum_{x \in \mathcal{X}} p_x s_r(np_x) \stackrel{(a)}{\leq} \sum_{r=2}^{\infty} \frac{\lambda^r}{n^{r-1}} \max_{t>0} \frac{s_r(t)}{r!}, \end{aligned} \quad (30)$$

where $s_r(t) \triangleq e^{-t}(1 - e^{-t})t^{r-1}$ and (a) follows because expectation is lesser than maximum. The following inequality, proved in the Appendix, relates the maximum values of $\frac{s_r(t)}{r!}$ and $\frac{s_{r-1}(t)}{(r-1)!}$:

$$\max_{t>0} \frac{s_r(t)}{r!} \leq \max_{t>0} \frac{s_{r-1}(t)}{(r-1)!}, \quad r \geq 3. \quad (31)$$

So, we have

$$\begin{aligned} L_{\mathbf{M}_0}(\lambda) &\leq \sum_{r=2}^{\infty} \frac{\lambda^r}{n^{r-1}} \max_{t>0} \frac{s_2(t)}{2} \\ &= \frac{\lambda^2 \gamma/n}{2(1-\lambda/n)}, \quad \lambda \in [0, n], \end{aligned} \quad (32)$$

where we have used $\max_{t>0} s_2(t) = \gamma$. This concludes the proof that \mathbf{M}_0 is sub-Gamma($\gamma/n, 1/n$). Since $\text{Var}(\mathbf{M}_0) \leq \frac{\gamma}{n}$, with equality when the underlying distribution is uniform over an alphabet of $n/1.4456$ letters, we have shown \mathbf{M}_0 to be sub-Gamma with the best (least) variance factor possible.

B. Sub-Gamma in multinomial sampling (Lemma 2, Part 2)

Using Bennett's inequality in (28) with $X = I(F_x = 0)$ and setting λ as λp_x , we get

$$L_{I(F_x=0)}(\lambda p_x) \leq (1-p_x)^n (1 - (1-p_x)^n) (e^{\lambda p_x} - \lambda p_x - 1). \quad (33)$$

Using the above in (27),

$$\begin{aligned} L_{\mathbf{M}_0}(\lambda) &\leq \sum_{x \in \mathcal{X}} (1-p_x)^n (1 - (1-p_x)^n) (e^{\lambda p_x} - \lambda p_x - 1) \\ &\stackrel{(a)}{\leq} \sum_{x \in \mathcal{X}} e^{-np_x} (1 - e^{-np_x} + 2p_x) \sum_{r=2}^{\infty} \frac{(\lambda p_x)^r}{r!} \\ &= \sum_{r=2}^{\infty} \frac{\lambda^r}{r! n^{r-1}} \sum_{x \in \mathcal{X}} p_x \left(s_r(np_x) + \frac{1}{n} u_r(np_x) \right) \\ &\leq \sum_{r=2}^{\infty} \frac{\lambda^r}{r! n^{r-1}} \left(\max_{t>0} s_r(t) + \frac{1}{n} \max_{t>0} u_r(t) \right), \end{aligned} \quad (34)$$

where $u_r(t) = 2e^{-t} t^r$ and (a) uses

$$e^{-np} - 2p \leq (1-p)^n \leq e^{-np}, \quad p \in [0, 1].$$

Now, since $\arg \max_{t>0} u_r(t) = r$ for all r , we see that

$$\max_{t>0} \frac{u_r(t)}{r!} \leq \max_{t>0} \frac{u_{r-1}(t)}{(r-1)!}.$$

So, we have

$$\begin{aligned} L_{\mathbf{M}_0}(\lambda) &\leq \left(\max_{t>0} \frac{s_2(t)}{2} + \frac{1}{n} \max_{t>0} \frac{u_2(t)}{2} \right) \sum_{r=2}^{\infty} \frac{\lambda^r}{n^{r-1}} \\ &\leq \frac{\lambda^2 (\gamma/n + \gamma'/n^2)}{2(1-\lambda/n)}, \quad \lambda \in [0, n], \end{aligned} \quad (35)$$

where we have used $\max_{t>0} u_2(t) = \gamma'$. This concludes the proof that M_0 is sub-Gamma($\gamma/n + \gamma'/n^2, 1/n$). Note that for $n \geq 5$, $\gamma/n + \gamma'/n^2 < 0.5/n$, where $0.5/n$ is the variance factor with which M_0 is sub-Gaussian on the right tail [12].

C. Counter example (Lemma 2, Part 3)

For $\alpha \in (0, n/2]$, let P be the uniform distribution on $\mathcal{X} = \{1, 2, \dots, n/\alpha\}$ with $p_x = \frac{\alpha}{n}$. It is easy to see that

$$\begin{aligned} L_{\mathbf{M}_0}(\lambda) &= \ln E[e^{\lambda \sum_{x \in \mathcal{X}} p_x (I(\mathbf{F}_x=0) - e^{-np_x})}] \\ &= \sum_{x \in \mathcal{X}} \ln E[e^{\frac{\lambda \alpha}{n} [I(\mathbf{F}_x=0) - e^{-\alpha}] }] \\ &= \frac{n}{\alpha} \ln \left(e^{\frac{\lambda \alpha}{n} (1-e^{-\alpha})} e^{-\alpha} + e^{\frac{\lambda \alpha}{n} (-e^{-\alpha})} (1-e^{-\alpha}) \right) \\ &\stackrel{(a)}{\geq} \frac{n}{\alpha} \left(\frac{\lambda \alpha}{n} (1-e^{-\alpha}) - \alpha \right) \\ &= \lambda(1-e^{-\alpha}) - n, \end{aligned} \quad (36)$$

where (a) follows by dropping the positive term $e^{\frac{\lambda \alpha}{n} (-e^{-\alpha})} (1-e^{-\alpha})$ in the argument of \ln . Set $\lambda = 2n$ in (36) to get

$$L_{\mathbf{M}_0}(2n) \geq n(1-2e^{-\alpha}). \quad (37)$$

For $\delta > 2e^{-n/2}$, we can choose a large enough $\alpha \in (0, n/2]$ to get

$$L_{\mathbf{M}_0}(2n) > n(1-\delta), \quad \delta > 0. \quad (38)$$

Now, if \mathbf{M}_0 is sub-Gaussian(c/n) with $c = 1/2 - \delta/2$, then

$$L_{\mathbf{M}_0}(2n) \leq \frac{(2n)^2(c/n)}{2} = n(1-\delta). \quad (39)$$

For any $\delta \in (0, 1]$, we can choose a large enough n followed by a large enough $\alpha \in (0, n/2]$ such that (38) holds. But (39) contradicts (38). So, \mathbf{M}_0 is not sub-Gaussian(c/n) for $c < 1/2$. This completes the proof of Lemma 2, Part 3.

D. $\mathbf{G}_k - \mathbf{M}_k$: Sub-Gaussian on left tail (Lemma 2, Part 4)

In [13], $\mathbf{G}_0 - \mathbf{M}_0$ was shown to be sub-Gamma($3E[\phi_1 + \phi_2 + \dots]/n^2, 1/n$) on the left tail. Here we prove that for $k \leq n-1$, $\mathbf{G}_k - \mathbf{M}_k$ is sub-Gaussian($\text{Var}(\mathbf{E}_k)$) on the left tail, resulting in (18) that improves (10) by a factor of $\sqrt{\ln n}$.

Recall the notation $\mathbf{E}_x = \frac{k+1}{n} I(\mathbf{F}_x = k+1) - p_x I(\mathbf{F}_x = k)$ and $\mathbf{E}_k = \mathbf{G}_k - \mathbf{M}_k$.

For $k \leq n-1$, we have $|\mathbf{E}_x| \leq 1$ and $E[\mathbf{E}_x] = 0$ for $x \in \mathcal{X}$. Applying Bennett's inequality (28) to \mathbf{E}_x , we get

$$L_{\mathbf{E}_x}(\lambda) \leq \text{Var}(\mathbf{E}_x) (e^\lambda - \lambda - 1). \quad (40)$$

Using the above in (22),

$$\begin{aligned} L_{\mathbf{E}_k}(\lambda) &\leq \sum_{x \in \mathcal{X}} \text{Var}(\mathbf{E}_x) (e^\lambda - \lambda - 1) \\ &= \text{Var}(\mathbf{E}_k) (e^\lambda - \lambda - 1). \end{aligned} \quad (41)$$

Therefore, for $k \leq n-1$, \mathbf{E}_k is sub-Poisson($\text{Var}(\mathbf{E}_k)$). From Lemma 1, for $k \leq n-1$, $-\mathbf{E}_k$ is sub-Gaussian($\text{Var}(\mathbf{E}_k)$).

This completes the proof of Lemma 2, Part 4.

E. $\mathbf{G}_k - \mathbf{M}_k$: Sub-Gamma on right tail (Lemma 2, Part 5)

Our proof generalizes the one in [13] for $k = 0$ to a general k . We start with (22) and proceed as follows.

$$\begin{aligned} L_{\mathbf{E}_k}(\lambda) &= \sum_{x \in \mathcal{X}} L_{\frac{n_{\mathbf{E}_x}}{k+1}} \left(\frac{k+1}{n} \lambda \right) \\ &\stackrel{(a)}{\leq} \sum_{x \in \mathcal{X}} \left(\frac{n}{k+1} \right)^2 \text{Var}(\mathbf{E}_x) \left(e^{\lambda \frac{k+1}{n}} - \lambda \frac{k+1}{n} - 1 \right) \\ &= \frac{\lambda^2}{2} \text{Var}(\mathbf{E}_k) \sum_{r=2}^{\infty} \left(\frac{k+1}{n} \right)^{r-2} \lambda^{r-2} \frac{2}{r!} \\ &\leq \frac{\lambda^2}{2} \text{Var}(\mathbf{E}_k) \sum_{r=2}^{\infty} \left(\frac{k+1}{n} \right)^{r-2} \left(\frac{\lambda}{3} \right)^{r-2} \\ &= \frac{\lambda^2 \text{Var}(\mathbf{E}_k)}{2(1-\lambda(\frac{k+1}{3n}))}, \quad 0 \leq \lambda < \frac{3n}{k+1}, \end{aligned} \quad (42)$$

where the manipulations to obtain (a) are similar to those in [13] and have been skipped. From (42), \mathbf{E}_k is sub-Gamma($\text{Var}(\mathbf{E}_k)$, $(k+1)/3n$). This completes the proof of Lemma 2, Part 5.

F. Confidence intervals (Theorem 5)

Since $-\mathbf{E}_k$ is sub-Gaussian($\text{Var}(\mathbf{E}_k)$), we have the following left tail bound for $k \leq n-1$:

$$\begin{aligned} &\text{With probability at least } 1-\delta, \\ &\mathbf{E}_k \geq -\sqrt{2\text{Var}(\mathbf{E}_k) \ln(1/\delta)}. \end{aligned} \quad (43)$$

Since \mathbf{E}_k is sub-Gamma($\text{Var}(\mathbf{E}_k)$, $\frac{k+1}{3n}$), we have the following right tail bound:

With probability at least $1-\delta$,

$$\mathbf{E}_k \leq \sqrt{2\text{Var}(\mathbf{E}_k) \ln(1/\delta)} + \frac{(k+1) \ln(1/\delta)}{3n}. \quad (44)$$

Let

$$\Phi = \frac{k+1}{k+2} \phi_{k+1} + \phi_{k+2}.$$

It can be shown that Φ is sub-Poisson($\text{Var}(\Phi)$), which follows by applying Bennett's inequality to $\frac{k+1}{k+2} I(\mathbf{F}_x=k+1) + I(\mathbf{F}_x=k+2)$ and we skip the details.

So, by Lemma 1, $-\Phi$ is sub-Gaussian($\text{Var}(\Phi)$). Since $\text{Var}(\Phi) \leq E[\Phi]$, we have that $-\Phi$ is sub-Gaussian($E[\Phi]$) as well. Therefore, we have the following left tail bound:

With probability at least $1-\delta$,

$$\begin{aligned} \Phi &\geq E[\Phi] - \sqrt{2 \ln(1/\delta) E[\Phi]} \\ &= \left(\sqrt{E[\Phi]} - \sqrt{0.5 \ln(1/\delta)} \right)^2 - 0.5 \ln(1/\delta), \\ \text{or } \sqrt{E[\Phi]} &\leq \sqrt{\Phi + 0.5 \ln(1/\delta)} + \sqrt{0.5 \ln(1/\delta)} \\ &\leq \sqrt{\Phi} + \sqrt{2 \ln(1/\delta)}, \end{aligned} \quad (45)$$

where the last inequality follows by using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$. From (24), $\text{Var}(\mathbf{E}_k) = \left(\frac{k+1}{n} \right)^2 E \left[\frac{k+2}{k+1} \Phi \right]$. Using this in (45) and rewriting, we have the following:

With probability at least $1-\delta$,

$$\begin{aligned} \sqrt{2\text{Var}(\mathbf{E}_k) \ln(1/\delta)} &\leq \frac{(k+1) \sqrt{2(\phi_{k+1} + \frac{k+2}{k+1} \phi_{k+2}) \ln(1/\delta)}}{n} \\ &\quad + \frac{2\sqrt{(k+1)(k+2) \ln(1/\delta)}}{n}. \end{aligned} \quad (46)$$

Theorem 5 results from (43), (44) and (46).

III. CONCLUSION AND FUTURE DIRECTION

We have improved the right tail concentration bounds of M_0 by showing that M_0 is sub-Gamma with its respective variance upper bounds as variance factors and $\frac{1}{n}$ as scale factor in both Poisson and multinomial sampling models. The variance factors are tighter than $\frac{0.5}{n}$ with which M_0 is sub-Gaussian under both Poisson and multinomial sampling models. In the case of Poisson sampling model, the variance factor with which M_0 is shown to be sub-Gamma is the best possible as it is exactly the variance of a uniform distribution over n/α for a suitable choice of α . We have also shown that M_0 , under Poisson sampling, cannot be sub-Gaussian with variance factor of $\frac{c}{n}$ for $c \in [0, 0.5)$. By exploiting the sub-Poisson nature of the Good-Turing estimator error, we improve upon previously established confidence intervals.

A natural future direction would be to establish if M_0 under the multinomial sampling model is sub-Gaussian with some $\frac{c}{n}$, where $c < 0.5$. Another interesting direction is to establish best-possible tail bounds for missing mass.

IV. APPENDIX: PROOF OF (31)

$$\begin{aligned} \frac{\max_{t>0} \frac{s_r(t)}{r!}}{\max_{t>0} \frac{s_{r-1}(t)}{(r-1)!}} &= \frac{1}{r} \frac{\max_{t>0} s_r(t)}{\max_{t>0} s_{r-1}(t)} \\ &\leq \frac{1}{r} \left(\frac{s_r(t)}{s_{r-1}(t)} \right) |_{t=\arg \max_{t>0} s_r(t)} \\ &= \frac{\arg \max_{t>0} s_r(t)}{r}. \end{aligned}$$

So, we need to show that $\arg \max_{t>0} s_r(t) \leq r$. We have

$$s'_r(t) = t^{r-2} e^{-t} [(r-1-t) + e^{-t} (2t-(r-1))].$$

As $(1-e^{-t})$ is increasing for $t > 0$ and $e^{-t} t^{r-1}$ is increasing for $t \in (0, r-1)$ (It has global maximum at $t = r-1$), we have that $s'_r(t) > 0$ for $t \in (0, r-1)$.

Also, $\frac{2t-(r-1)}{e^t} \leq \frac{2t}{e^t} \leq 2e^{-1} < 1$ and $(r-1)-t < -1$, for $t \in (r, \infty)$. So, we have $s'_r(t) < 0$ for $t \in (r, \infty)$.

Since $s'_r(t)$ is evidently continuous and we have $s'_r(t) > 0$ for $t \in (0, r-1)$ and $s'_r(t) < 0$ for $t \in (r, \infty)$, there exists t with $s'_r(t) = 0$ only for $t \in [r-1, r]$. So, we have $\arg \max_{t>0} s_r(t) \leq r$.

REFERENCES

- [1] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. 237–264, 1953.
- [2] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," in *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ser. ACL '96, 1996, pp. 310–318.
- [3] D. A. McAllester and R. E. Schapire, "On the convergence rate of good-turing estimators," in *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, 2000, pp. 1–6.
- [4] A. Chao and S.-M. Lee, "Estimating the number of classes via sample coverage," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.
- [5] W. A. Gale and G. Sampson, "Good-Turing frequency estimation without tears," *Journal of Quantitative Linguistics*, vol. 2, no. 3, pp. 217–237, 1995.
- [6] A. Orlitsky, N. Santhanam, and J. Zhang, "Always good-turing: Asymptotically optimal probability estimation," in *Annual Symposium on Foundations of Computer Science - Proceedings*, 2003, pp. 179–188.
- [7] E. Drukh and Y. Mansour, "Concentration bounds for unigram language models," *J. Mach. Learn. Res.*, vol. 6, pp. 1231–1264, Dec. 2005.
- [8] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong consistency of the good-turing estimator," in *2006 IEEE International Symposium on Information Theory*, July 2006, pp. 2526–2530.
- [9] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Optimal probability estimation with applications to prediction and classification," in *Proceedings of the 26th Annual Conference on Learning Theory*, vol. 30, 2013, pp. 764–796.
- [10] A. Orlitsky and A. T. Suresh, "Competitive distribution estimation: Why is good-turing good," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2143–2151.
- [11] D. McAllester and L. Ortiz, "Concentration inequalities for the missing mass and for histogram rule error," *J. Mach. Learn. Res.*, vol. 4, pp. 895–911, Dec. 2003.
- [12] D. Berend and A. Kontorovich, "On the concentration of the missing mass," *Electron. Commun. Probab.*, vol. 18, p. 7 pp., 2013.
- [13] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian, "Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications," *Bernoulli*, vol. 23, no. 1, pp. 249–287, 02 2017.
- [14] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press, 2013.
- [15] D. Dubhashi and D. Ranjan, "Balls and bins: A study in negative dependence," *Random Structures & Algorithms*, vol. 13, no. 2, pp. 99–124, Dec. 1998.

Differential Phase Encoding Scheme for Measurement-Device-Independent Quantum Key Distribution

Shashank Kumar Ranu

*Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India
ee16s300@ee.iitm.ac.in*

Anil Prabhakar

*Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India
anilpr@ee.iitm.ac.in*

Prabha Mandayam

*Department of Physics
Indian Institute of Technology Madras
Chennai, India
prabhamd@physics.iitm.ac.in*

Abstract—This paper proposes a measurement-device-independent quantum key distribution (MDI-QKD) scheme based on differential phase encoding. The differential phase shift MDI-QKD (DPS-MDI-QKD) couples the advantages of DPS-QKD with that of MDI-QKD. The proposed scheme offers resistance against photon number splitting attack and phase fluctuations as well as immunity against detector side-channel vulnerabilities. The design proposed in this paper uses weak coherent pulses in a superposition of three orthogonal states, corresponding to one of three distinct paths in a delay-line interferometer. The classical bit information is encoded in the phase difference between pulses traversing successive paths. This 3-pulse superposition offers enhanced security compared to using a train of pulses by decreasing the learning rate of an eavesdropper and unmasking her presence with an increased error rate upon application of intercept and resend attack and beamsplitter attack. The proposed scheme employs phase locking of the sources of the two trusted parties so as to maintain the coherence between their optical signal, and uses a beamsplitter (BS) at the untrusted node (Charlie) to extract the key information from the phase encoded signals.

Index Terms—MDI-QKD, DPS-QKD, photon number splitting attack, intercept and resend attack, beamsplitter attack, phase fluctuations, secure key rate.

I. INTRODUCTION

Though unconditional security of QKD has been theoretically proven, practical implementations of QKD protocols contain imperfections which compromise their security. For example, use of a weak coherent sources instead of a single photon sources make QKD implementations vulnerable to photon number splitting attacks. Similarly, attacks exploiting imperfections in detectors have posed a serious threat to the practical security of QKD systems. Taking advantage of the side channel information from the detectors, attacks such as time-shift attack [1], phase-remapping attack [2], detector control attack [3]–[6] and detector dead-time attack [7], which render practical QKD systems insecure, have been demonstrated. Device independent QKD [8], [9] was introduced as a solution to detector side-channel attacks. However, this requires detectors with near unity detection efficiency and even then has an extremely low key rate, thereby making it practically infeasible. Measurement-Device-Independent

Quantum Key Distribution (MDI-QKD), introduced by Lo, Curty, and Qi [10], offers a practical solution to attacks which exploit detector imperfections. The security proof of MDI-QKD follows along the lines of time reversal EPR based QKD, without making any assumptions on the measurements or measuring devices, hence making it measurement device independent. In MDI-QKD Alice and Bob prepare quantum states independently and send it to a measurement node (Charles/Eve) through the quantum channel. It is assumed that the measurement node is under the control of an eavesdropper (Eve). Charles makes the measurement and announces the result. Based upon Charles announcement, Alice and Bob carry out key reconciliation. The protocol implicitly assumes that the measurement device is under Eves control, thereby circumventing all the detector side-channel issues.

MDI-QKD based on polarization encoding requires polarization maintaining fiber for its implementation. This along with the time-dependent nature of the birefringence effect in optical fiber makes implementation of polarization based MDI-QKD a challenging task. Hence, MDI-QKD protocols employing phase encoding have been proposed [11]. However, random phase fluctuations still pose a challenge to the implementation of such phase encoded QKD protocols. DPS-QKD was introduced by Inoue et.al. [12], [13] as a solution to this problem of phase fluctuations. Here, information is encoded in the phase difference of two consecutive pulses. Any two consecutive pulses experience nearly similar phase and polarization changes along the optical fiber channel, hence the effects of phase and polarization fluctuations cancel out.

In this paper, we propose an MDI-QKD scheme which attempts to encapsulate the advantages of differential phase shift QKD. We use the variant of DPS wherein Alice and Bob transmit photons in a linear superposition of three states [12]. This enhances the security of our proposed design against individual attacks. Section II of this paper describes the proposed MDI protocol. Section III presents the simulation results pertaining to secure key rate generation and its dependence on the quantum channel length. We conclude the paper in section IV.

II. DIFFERENTIAL PHASE ENCODING SCHEME

In this section, we present the design of MDI protocol based on differential phase encoding. Fig. 1 shows the setup of the proposed protocol. Alice and Bob use weak coherent

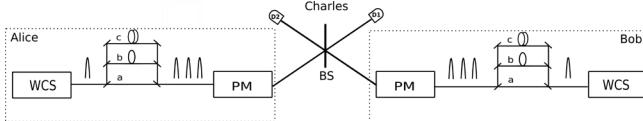


Fig. 1. Schematic of differential phase encoded MDI-QKD.

sources to obtain mean photon numbers of less than one per pulse. Couplers and beamsplitters are used to convert the weak coherent pulses into a linear superposition of three states. The phase modulator, driven by a quantum random number generator (QRNG), applies a phase of 0 or π to the pulses that come out of the delay lines. Alice and Bob send these randomly phase modulated weak coherent pulses to the measurement node (Charles). Charles uses a beamsplitter to perform the measurement. The input state to the beam splitter may be expressed as,

$$|\psi_{in}\rangle = \frac{1}{3} (|\pm\alpha\rangle_a + |\pm\alpha\rangle_b + |\pm\alpha\rangle_c)_{Alice} \otimes (|\pm\alpha\rangle_a + |\pm\alpha\rangle_b + |\pm\alpha\rangle_c)_{Bob}.$$

Here, we assume a mean photon of $|\alpha^2|$ before the delay line. The beamsplitter coefficients are chosen such that a photon has an equal probability of traversing through each path. This translates to mean photon number of $|\frac{\alpha^2}{3}|$ in an individual path of the delay line.

Fig.2 shows a typical 50:50 beamsplitter. We assume that weak coherent pulses fall on both the input ports the beamsplitter. Hence, we write input state as,

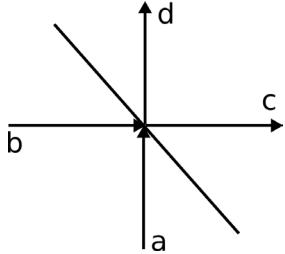


Fig. 2. a and b are the input ports of beamsplitter. The output ports are c and d.

$$|\psi\rangle_{in} = |\alpha\rangle_a \otimes |\beta\rangle_b. \quad (1)$$

In terms of displacement operator, the input to the beamsplitter can be written as,

$$|\psi\rangle_{in} = D(\alpha)|0\rangle_a D(\beta)|0\rangle_b. \quad (2)$$

After applying the beamsplitter transformation, we get

$$|\psi\rangle_{out} = D\left(\frac{\alpha+\beta}{\sqrt{2}}\right)|0\rangle_c D\left(\frac{\alpha-\beta}{\sqrt{2}}\right)|0\rangle_d. \quad (3)$$

The above equation can be simplified to

$$|\psi\rangle_{out} = \left| \frac{\alpha+\beta}{\sqrt{2}} \right\rangle_c \left| \frac{\alpha-\beta}{\sqrt{2}} \right\rangle_d. \quad (4)$$

In case of the MDI protocols, $|\alpha\rangle = |\beta\rangle$. Using this in eq (4), we tabulate the beamsplitter outputs for different phases between Alice and Bob's weak coherent pulses.

Table I : Beamsplitter output for different input configurations

Input to the beamsplitter	Beamsplitter output
$ \alpha\rangle_a \alpha\rangle_b$	$ \sqrt{2}\alpha\rangle_c 0\rangle_d$
$ \alpha\rangle_a -\alpha\rangle_b$	$ 0\rangle_c \sqrt{2}\alpha\rangle_d$
$ -\alpha\rangle_a \alpha\rangle_b$	$ 0\rangle_c -\sqrt{2}\alpha\rangle_d$
$ -\alpha\rangle_a -\alpha\rangle_b$	$ - \sqrt{2}\alpha\rangle_c 0\rangle_d$

We observe from the Table I that photons come out of port c (d) when the phase difference between the two input pulses is 0 (π). We make use of this fact in designing our protocol.

Photons from Alice and Bob contribute to the key only when they take the same path through their respective delay lines. This implies that unlike DPS-QKD, the key is no longer encoded in the phase difference of two consecutive pulses of Alice or Bob; rather, the key information is now in the phase difference between the corresponding time-bins of Alice and Bob. Charles uses the beamsplitter to determine whether Alice and Bob applied the same phase ((0, 0) or (π , π)) or different phase ((0, π) or (π , 0)). Detector D1 (D2) clicks when Alice and Bob have applied the same (opposite) phase. For each time slot, the intermediate node (Charles) announces which detector clicks. If the result is inconclusive, Charles announces "?". Key reconciliation is thus carried out as:

- Outcome = D1 \Rightarrow Alice and Bob do nothing.
- Outcome = D2 \Rightarrow Alice/Bob performs bit flip.
- Outcome = ? \Rightarrow Alice and Bob discard the bit.

Though the setup shown in Fig. 1 generates secure keys in an ideal scenario, practical implementation of the proposed scheme requires modifications (see Fig. 3). Key generation requires detection of two time synchronized photons by a single detector. This is practically infeasible due to the finite dead-time of single-photon detectors. The acousto-optic deflectors (AOD) and two additional single-photon detectors are added to the setup to alleviate this problem. Acousto-optic deflector is used to route the two photons of each time-bin to two different single-photon detectors. This results in a slight modification to key reconciliation step. Now, Charles announces which pair of detectors clicked for each time slot. Alice and Bob keep their bits unchanged when $D1$ and $D2$ click. Clicking of $D3$ and $D4$ forces one of Alice or Bob to perform a bit flip operation. Any other combination of detector clickings contributes to inconclusive measurement. Alice and Bob use independent laser sources in this scheme. They need to have a common phase reference as the information is encoded in the phase difference between the corresponding time-bins of Alice and Bob. Optical phase-locked loop (OPLL) is a suitable technique for locking Alice's and Bob's source so as to maintain coherence between their optical signals. OPLL

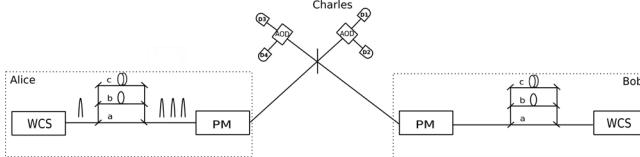


Fig. 3. Practical implementation of DPS-QKD.

technique is widely used in coherent detection for increasing the receiver's sensitivity [14], [15]. This phase-locking technique offers advantages such as a simple realization set-up and requires only off-the-shelf optical components [16].

This 3-pulse encoding used above is a variant of the pulse-train DPS-QKD protocol [13] which in turn is based on the B92 protocol [17]. In the pulse-train DPS protocol, Alice generates a train of pulses and applies a phase of 0 or π to the pulses randomly. This phase modulated pulse train is sent to Bob, who passes the incoming pulses through an unbalanced Mach-Zehnder interferometer (MZI). Depending upon the phase difference between two successive pulses, constructive or destructive interference happens. A measurement device independent QKD protocol based on this has been proposed recently [18].

Our MDI protocol based on the 3-pulse encoding offers a low key rate compared to the scenario when Alice and Bob use a train of pulses as their signal. This happens due to the fact that only photons taking the same path in the delay lines of Alice and Bob contribute to the key. The probability of a photon taking same path in both Alice's and Bob's side is $\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{1}{3}$, thereby making the sifted key rate $\frac{1}{3}$. However, this scheme offers enhanced security against individual attacks than its counterpart.

This paper assumes that Eve is capable of carrying out only intercept and resend attack and beamsplitter attack at either Alice's or Bob's end. For the intercept and resend attack, Eve's setup consists of an unbalanced Mach-Zehnder interferometer and two single-photon detectors to retrieve the differentially encoded phase information. She detects photons at four possible time-bins (a,b,c and d). When Eve detects a photon in a and d time-bins, she gains no phase information. Detection in b and c time-bins provide her partial phase information. The information is partial in the sense that she gains knowledge regarding the phase difference between the two neighbouring time-bins but lacks information about individual phase of those two time-bins. This implies that she cannot regenerate a photon in a superposition of three time-bins having the same phase as that of Alice's pulses with absolute certainty. Suppose Eve detects a photon at time-bin a. This detection event doesn't provide any information about Δb (phase difference between time-bins a and b) and Δc (phase difference between time-bins b and c). So, she applies the phases in the three time-bins randomly. With one-fourth of probability, she guesses Δb and Δc correctly. But 75% of the time she makes mistakes in guessing Δb and Δc , thereby inducing error in the key. The errors induced in the key for all the possible scenarios

due to this attack have been shown in Table II. Any attempt of eavesdropping using intercept and resend attack introduces error in 33% of the sifted key in 3-pulse DPS-QKD [19] while error creeps in only 25% of the sifted key in pulse train DPS-MDI [20]. Intercept and resend attack reveals 33% of the sifted key to an eavesdropper in 3-pulse DPS-MDI-QKD compared to 50% key learning rate in pulse train DPS-MDI, thereby making 3-pulse DPS more secure.

The probability of getting one bit of sifted key information using beamsplitter attack also reduces by 25% in our implementation as compared to using a train of pulses for differential encoding [19] [20].

III. SECURE KEY RATE OF PROPOSED DESIGN

In this section, we use the sifted key rate and Eve's learning rate to obtain an expression for the secure key rate of our DPS-MDI-QKD protocol. Sifting of the key follows the raw key generation in any QKD protocol. Eavesdropping and imperfect devices used in the implementation of the protocol induce errors in the sifted key. Alice and Bob carry out error correction of this erroneous sifted key. However, this error correction is performed over a classical communication channel which is insecure. This leads to the leaking of information to the eavesdropper. Hence, privacy amplification follows this error correction step so as to reduce the eavesdropper's information below a certain threshold.

As per Shannon's noiseless coding theorem, the minimum number of bits (k) that Alice and Bob need to exchange so as to remove the errors from their shared bit string of length n is given as,

$$\lim_{n \rightarrow \infty} \frac{k}{n} = -e \log_2 e - (1-e) \log_2 (1-e) \equiv h(e), \quad (5)$$

where e is the Quantum Bit Error Rate (QBER). Privacy amplification follows the error correction step in the classical post-processing of any QKD protocol. The main aim of privacy amplification is to figure out the shrinking factor (τ) by which error-corrected key needs to be compressed so as to limit the amount of Eve's information about the key to less than some specified threshold. The shrinking factor (τ) is calculated by the method of generalized privacy amplification theory [21] as per which length of the final key should be,

$$r = n\tau - k - t. \quad (6)$$

Here, n is the length of the sifted key, k is the number of bits disclosed during error correction, t is a security parameter and τ is the shrinking factor obtained as,

$$\tau = \frac{-\log_2 p_c}{n}, \quad (7)$$

where, p_c is the average collision probability. The key step in the security proof of any QKD protocol is to find the bounds on p_c for specific types of attacks.

If N is the length of the transmission, then $n = NR_{\text{sifted}}$. Secure key generation rate is then defined as,

Table II : Error rate due to intercept and resend attack in 3-pulse DPS-QKD

Eve's detection time-bin	Eve sends Δb and Δc	Error free bits known to Eve	Error free bits known to Eve
<i>a</i> with $\frac{1}{6}$ probability (Eve gains no information)	P (both correct)= $\frac{1}{4}$	No error is induced	-
	P (both correct)= $\frac{1}{4}$	$\frac{1}{6} \times \frac{1}{4} \times \frac{2}{3} = \frac{1}{36}$	-
	P (Δb correct, Δc wrong)= $\frac{1}{4}$	$\frac{1}{6} \times \frac{1}{4} \times \frac{1}{3} = \frac{1}{72}$	-
	P (Δb wrong, Δc correct)= $\frac{1}{4}$	$\frac{1}{6} \times \frac{1}{4} \times \frac{1}{3} = \frac{1}{72}$	-
<i>b</i> with $\frac{1}{3}$ probability (Eve learns Δb)	P (Δc wrong)= $\frac{1}{2}$	$\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{18}$	$\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{18}$
	P (Δc correct)= $\frac{1}{2}$	No error is introduced	$\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{18}$
<i>c</i> with $\frac{1}{3}$ probability (Eve learns Δc)	P (Δb wrong)= $\frac{1}{2}$	$\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{18}$	$\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{18}$
	P (Δb correct)= $\frac{1}{2}$	No error is introduced	$\frac{1}{3} \times \frac{1}{2} \times \frac{1}{3} = \frac{1}{18}$
<i>d</i> with $\frac{1}{6}$ probability (Eve gains no information)	P (both correct)= $\frac{1}{4}$	No error is induced	-
	P (both correct)= $\frac{1}{4}$	$\frac{1}{6} \times \frac{1}{4} \times \frac{2}{3} = \frac{1}{36}$	-
	P (Δb correct, Δc wrong)= $\frac{1}{4}$	$\frac{1}{6} \times \frac{1}{4} \times \frac{1}{3} = \frac{1}{72}$	-
	P (Δb wrong, Δc correct)= $\frac{1}{4}$	$\frac{1}{6} \times \frac{1}{4} \times \frac{1}{3} = \frac{1}{72}$	-

$$\begin{aligned}
 R &= \lim_{N \rightarrow \infty} \frac{r}{N} = \frac{n\tau - k - t}{N} \\
 &= \frac{NR_{\text{sifted}}\tau - k - t}{N} \\
 &= \lim_{n \rightarrow \infty} R_{\text{sifted}}(\tau - \frac{k}{n} - \frac{t}{n}). \quad (8)
 \end{aligned}$$

As $n \rightarrow \infty$, $\frac{t}{n} = 0$

Practical algorithms do not work at Shannon's limit. Defining $f(e)$ as the ratio of the practical algorithm's performance and Shannon's limit, Eq. (5) and (8) can be rewritten as

$$\lim_{n \rightarrow \infty} \frac{k}{n} = f(e)h(e). \quad (9)$$

$$R = R_{\text{sifted}}\{\tau + f(e)[e \log_2 e + (1 - e) \log_2(1 - e)]\}. \quad (10)$$

Estimation of the secure key rate in 3-pulse DPS-MDI-QKD using Eq. (10) requires knowledge of parameters such as the error rate (e) and sifted key rate. The sifted key rate is calculated as $s\gamma p_{\text{click}}$, where s is the sifting parameter and γ is the repetition rate of transmission. p_{click} is the probability that Bob detects a photon in a given clock cycle and is given by,

$$p_{\text{click}} = p_{\text{signal}} + p_{\text{dark}}, \quad (11)$$

where, p_{signal} is the probability of clicking due to signal and p_{dark} is the probability of clicking due to dark count. The probability of clicking due to the signal depends on transmission efficiency (T) of the channel and mean photon number per pulse (μ) as shown below:

$$p_{\text{signal}} = \mu T. \quad (12)$$

The transmission efficiency of the quantum channel can be expressed as,

$$T = \eta 10^{\frac{-(\alpha L + L_s)}{10}}, \quad (13)$$

where η is the detector efficiency, α is the loss of the channel in dB/Km, L is the length of the channel and L_s is the loss of detector's setup.

Eq(14) gives the probability of clicking due to dark counts as,

$$p_{\text{dark}} = 2d. \quad (14)$$

The error rate used in Eq. (10) has contributions from both the signal as well as the dark count and hence, can be expressed as,

$$e = \frac{\frac{1}{2}p_{\text{dark}} + bp_{\text{signal}}}{p_{\text{signal}}}, \quad (15)$$

where b is the baseline error of the QKD system. Dark counts are random in nature which makes half of the dark count clicks correct. Hence a factor of $\frac{1}{2}$ is included in Eq. (15).

Here, it is assumed that Eve is capable of employing only intercept and resend, and beamsplitter attack. The collision probability (p_{c0}) is set to 1 for bits known to Eve and is $\frac{1}{2}$ for bits unknown to Eve. When Eve employs intercept and resend attack, an error is introduced in 33% of the key and she learns about 33% of the key [19]. Given that the error rate is e , Eve can successfully apply intercept and resend attack on $3e$ bits and can learn about e of them. Similarly, application of beamsplitter attack gives her information about $\frac{3}{4}\mu(1-T)$ bits [19]. Hence, $p_{c0} = \frac{1}{2}$ for $1 - \frac{3}{4}\mu(1-T) - e$ fraction of bits and 1 for the remaining bits.

Knowledge of average collision probability for an n -bit string is required to find out the shrinking factor for the 3-pulse-DPS-MDI-QKD. For an n -bit string, the average collision probability is given as,

$$p_c = p_{c0}^n. \quad (16)$$

Thus, the shrinking factor can be calculated as ,

$$\tau = \frac{-\log_2 p_c}{n} \quad (17)$$

$$= -\log_2 p_{c0} \quad (18)$$

$$= 1 - \frac{3}{4}\mu(1-T) - e. \quad (18)$$

Thus, secure key rate for 3-pulse DPS-MDI-QKD can be written as,

$$R_{\text{DPS-QKD}} = s\gamma p_{\text{click}} \{ \tau + f(e) [e \log_2 e + (1-e) \log_2(1-e)] \}, \quad (19)$$

where τ , p_{click} and e are obtained as shown in Eq. (18), (11) and (15) respectively.

Fig. 4 shows the effect of channel length on the secure key rate. The values of all the fixed parameters of the system are taken from the reference [20]. The mean photon number is set to 0.2. The values of the parameters used in this simulation are $\gamma = 10\text{MHz}$, $\alpha = 0.2 \text{ dB/km}$, $b = 0.01$, $L_s = 1\text{dB}$ and $\eta = 10\%$.

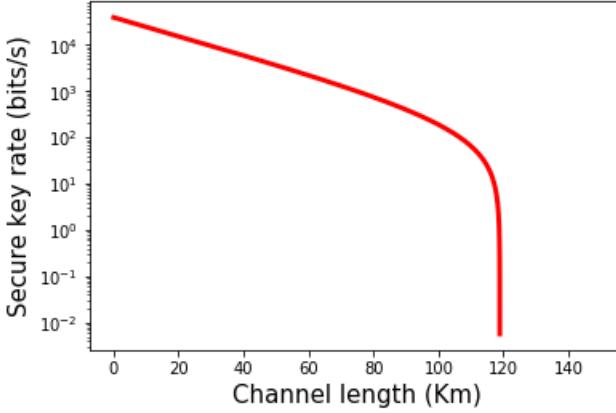


Fig. 4. Secure key rate as a function of channel length for 3-pulse DPS-MDI

IV. CONCLUSIONS

We have presented a novel MDI-QKD scheme where key information is encoded between pulses generated at corresponding time-bins of Alice and Bob. We report its superior performance against intercept and resend and beamsplitter attack, as compared to other differential phase encoded MDI schemes proposed in the literature. We have also carried out the key rate analysis of the proposed scheme and numerically simulated the variation of key rate with channel length.

The protocol presented here employs weak coherent sources for key generation. It would be interesting to check its functionality for ideal single-photon source. Finite-key analysis of the presented protocol is another interesting research problem. This analysis would involve mapping the protocol to an equivalent entanglement based protocol.

One interesting open problem is to examine whether the measurement at the untrusted node (Charles/Eve) may be carried out via a balanced Mach-Zehnder Interferometer (MZI) rather than a beamsplitter. Thus far, all phase-based MDI schemes in the literature have used only a beamsplitter at Charles' end. It is possible that using an MZI might provide information about the differential phase between Alice and Bob's pulses more frequently than a beamsplitter which relies on photons traversing identical paths at both sources, and hence lead to a higher sifted key rate. However, such a protocol would entail a more involved reconciliation step to obtain the sifted key. Identifying a suitable MZI-based scheme for MDI-QKD is thus an interesting avenue for future work.

REFERENCES

- [1] Yi Zhao, Chi-Hang Fred Fung, Bing Qi, Christine Chen, and Hoi-Kwong Lo. Quantum hacking: Experimental demonstration of time-shift attack against practical quantum-key-distribution systems. *Physical Review A*, 78(4):042333, 2008.
- [2] Feihu Xu, Bing Qi, and Hoi-Kwong Lo. Experimental demonstration of phase-remapping attack in a practical quantum key distribution system. *New Journal of Physics*, 12(11):113026, 2010.
- [3] Lars Lydersen, Carlos Wiechers, Christoffer Wittmann, Dominique Elser, Johannes Skaar, and Vadim Makarov. Hacking commercial quantum cryptography systems by tailored bright illumination. *Nature photonics*, 4(10):686, 2010.
- [4] ZL Yuan, JF Dynes, and AJ Shields. Resilience of gated avalanche photodiodes against bright illumination attacks in quantum cryptography. *Applied physics letters*, 98(23):231104, 2011.
- [5] ZL Yuan, JF Dynes, and AJ Shields. Response to comment on resilience of gated avalanche photodiodes against bright illumination attacks in quantum cryptography[appl. phys. lett. 99, 196101 (2011)]. *Applied physics letters*, 99(19):196101, 2011.
- [6] Ilja Gerhardt, Qin Liu, Antia Lamas-Linares, Johannes Skaar, Christian Kurtsiefer, and Vadim Makarov. Full-field implementation of a perfect eavesdropper on a quantum cryptography system. *Nature communications*, 2:349, 2011.
- [7] Henning Weier, Harald Krauss, Markus Rau, Martin Fürst, Sebastian Nauerth, and Harald Weinfurter. Quantum eavesdropping without interception: an attack exploiting the dead time of single-photon detectors. *New Journal of Physics*, 13(7):073024, 2011.
- [8] Dominic Mayers and Andrew Yao. Self testing quantum apparatus. *arXiv preprint quant-ph/0307205*, 2003.
- [9] Antonio Acín, Nicolas Brunner, Nicolas Gisin, Serge Massar, Stefano Pironio, and Valerio Scarani. Device-independent security of quantum cryptography against collective attacks. *Physical Review Letters*, 98(23):230501, 2007.
- [10] Hoi-Kwong Lo, Marcos Curty, and Bing Qi. Measurement-device-independent quantum key distribution. *Physical review letters*, 108(13):130503, 2012.
- [11] Kiyoshi Tamaki, Hoi-Kwong Lo, Chi-Hang Fred Fung, and Bing Qi. Phase encoding schemes for measurement-device-independent quantum key distribution with basis-dependent flaw. *Physical Review A*, 85(4):042307, 2012.
- [12] Kyo Inoue, Edo Waks, and Yoshihisa Yamamoto. Differential phase shift quantum key distribution. *Physical Review Letters*, 89(3):037902, 2002.
- [13] K Inoue, E Waks, and Y Yamamoto. Differential-phase-shift quantum key distribution using coherent light. *Physical Review A*, 68(2):022317, 2003.
- [14] JM Kahn, BL Kasper, and KJ Pollock. Optical phaselock receiver with multigigahertz signal bandwidth. *Electronics Letters*, 25(10):626–628, 1989.
- [15] Leonid G Kazovsky. Decision-driven phase-locked loop for optical homodyne receivers: Performance analysis and laser linewidth requirements. *IEEE Transactions on Electron devices*, 32(12):2630–2639, 1985.
- [16] Valter Ferrero and S Camatel. Optical phase locking techniques: an overview and a novel method based on single side sub-carrier modulation. *Optics express*, 16(2):818–828, 2008.
- [17] Charles H Bennett. Quantum cryptography using any two nonorthogonal states. *Physical review letters*, 68(21):3121, 1992.
- [18] Agnes Ferenczi. Security proof methods for quantum key distribution protocols. 2013.
- [19] S. K. Ranu, G. K. Shaw, A. Prabhakar, and P Mandayam. Security with 3-pulse differential phase shift quantum key distribution. *IEEE WRAP*, 2017.
- [20] Eleni Diamanti. *Security and implementation of differential phase shift quantum key distribution systems*. Stanford University, 2006.
- [21] Charles H Bennett, Gilles Brassard, Claude Crépeau, and Ueli M Maurer. Generalized privacy amplification. *IEEE Transactions on Information Theory*, 41(6):1915–1923, 1995.

Truthful Double Auction Based VM Allocation for Revenue-Energy Trade-Off in Cloud Data Centers

Yashwant Singh Patel*, Animesh Nighojkar†, Rajiv Misra*

*Department of Computer Science and Engineering

Indian Institute of Technology Patna, India - 801106

†Department of Computer Science and Engineering

Medi-Caps Institute of Technology and Management, India - 453331

E-mail:{yashwant.pcs17, rajivm}@iitp.ac.in, nighojkaranimesh@gmail.com

Abstract—With the advances in virtualization technologies, cloud has emerged as a flexible and cost-effective service paradigm by provisioning on-demand VM resources to users via a pay-per-use business model. In cloud data centers, effective resource provisioning is required with the aim of minimizing energy consumption and maximizing cloud provider's revenue. However, the existing mechanisms have either focused on the optimization of energy, or the profit of cloud service provider (CSP) while incurring inefficient resource allocation. Thus to address these fundamental research challenges and to balance the trade-off between energy and revenue, we propose a Vickrey-Clarke-Groves (VCG) based truthful double auction mechanism (TDAM). In this paper, first, we have formulated a joint optimization problem and prove it NP-hard by reducing it to a multi-dimensional bin-packing problem. Then we design TDAM, a truthful double auction scheme and propose an efficient winning bid algorithm for VM allocation and a VCG based mechanism for calculating payment of each bid. Being a double auction, TDAM allows both the buyers (VMs) and the sellers (PMs) to submit their bids and asks respectively, and performs allocation based on the energy consumption, while upholding truthfulness, in order to avoid falsification of the submitted bid or ask values. Through theoretical analysis and extensive experiments we show that the TDAM makes a significant contribution while maintaining truthfulness, individual rationality, economic efficiency, and has polynomial time complexity.

Index Terms—Virtual machine (VM), Physical machine (PM), VCG auction, Asymptotic approximation ratio, Incentive mechanism, Truthfulness

I. INTRODUCTION

Cloud computing provides flexible and cost-effective services by enabling on-demand provisioning of computational resources based on the pay-per-use business model. With the help of cloud platforms such as Amazon EC2 and Microsoft Azure, individual users can submit their request of required resources (e.g. CPU, memory, network bandwidth, and storage) to cloud service providers (CSPs). The CSPs then make resources available to users in the form of VMs in exchange for financial remuneration [1]. An efficient VM allocation is a challenging problem because while satisfying various user requirements, it has to maintain a trade-off between CSP's profit and energy cost minimization. In order to maximize the revenue, a CSP will always try to allocate as many VMs as possible which consequently increase the power consumption in terms of the number of active physical machines (PMs). Due to heterogeneity in the number of resources, an inefficient resource allocation may result in more number of PMs with a tremendous increase in energy consumption. Thus there must be a trade between the VMs requesting the cloud resources and PMs which are providing the resources. The auction is become one of the well-known trading forms as it allocates the resources of sellers to buyers and allows competitive price discovery as well as maintains efficient and fair resource allocation[2]. In this work, we propose a truthful double auction based

mechanism (TDAM) for VM allocation. The key contributions of this work can be listed as follows: Firstly, the VM allocation problem of maximizing revenue and minimizing energy cost in cloud data centers is formulated as an NP-Hard problem then we design an efficient *truthful* double auction mechanism *TDAM* to solve it. Under this scheme, we have mainly proposed winning bid determination algorithm by using the maximum matching algorithm and then VCG based payment strategy for the winning bids. Secondly, with the help of theoretical analysis and simulations, it is shown that the TDAM follows truthfulness, individual rationality, and economic efficiency with polynomial time complexity. Finally, we have validated the desirable auction properties through extensive simulation.

II. RELATED WORK

This section categorizes the related work study into two parts: The first part belongs to the related auction approaches and the second part is dedicated to VM allocations from cloud systems literature. Auction theory in different fields has been widely studied, the existing auction schemes cannot satisfy all the four fundamental properties elaborated in previous section. The key challenges for any auction design are the heterogeneity of items and optimal allocation of resources to achieve the truthfulness. In Table 1, we have shown the study of related works along with their major differences. Certainly, such auction schemes are not feasible as they all assume features specifically concern with the studied problems and cannot capture the individual aspects and realistic issues of the VM allocation problem. There are

Reference	Truthful	Double Auction	Heterogeneous Item
[9]	–	✗	✓
[10]	✓	✗	✓
[11]	–	✗	✓
[13]	✓	✓	✗
[14]	✗	✓	✗

Table I
EXISTING AUCTION MECHANISMS

a very less studies on the auction mechanisms for VM allocation in cloud literature. In [5][6] two auction schemes are designed for sharing of cloudlet resources in the mobile cloud environment. However, many of these auction approaches are specifically designed for homogeneous task model and limit the auctioning in a one-to-one matching. [4] proposed two different truthful auction approaches for mobile device clouds. The authors have designed a VCG-based auction scheme for homogeneous task model and winning bids determination algorithm for heterogeneous task model. Wang et al. [8] discussed the problem of an efficient VM allocation through solving the problem of energy minimization. However, the existing works are far from the efficient pricing mechanisms for energy

cost minimization and maximum revenue achieved from the users' requests.

III. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a data center having κ physical machines (PMs) or servers to cater to the requirements of η virtual machines (VMs). A VM can be allocated to only one server. It cannot work on different resources from different servers. Also, the resource requirements of a VM cannot be compromised upon. To solve this allocation problem while maximizing the revenue from VM allocation and minimizing the total energy consumption of the servers at the data center, we propose a VCG-based double auction mechanism for allocation and payment. Our model has three phases: (i) Auction Setup: We take the buyer and seller details as input and arrange them accordingly. (ii) Winner Determination and Allocation: We allocate the servers as per their asks and energy consumption values to the VMs as per their bids and (iii) Payment: Using a mechanism similar to the VCG based payment scheme, we calculate the payment the VMs need to give and the payment the servers would get.

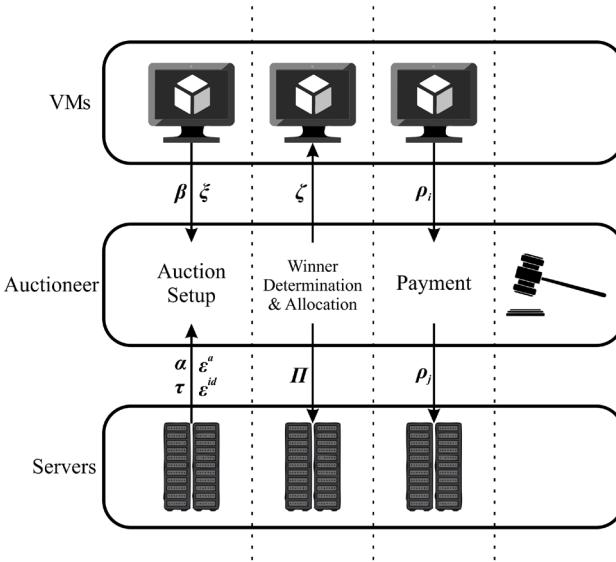


Figure 1. Auction Model

A. The TDAM Model

We consider a set of η VMs $\Theta = \{B_1, B_2, B_3, \dots, B_\eta\}$ requesting resources from among a set of κ servers $\Gamma = \{S_1, S_2, S_3, \dots, S_\kappa\}$ at a single data center. We assume only one type of resource, CPU, for ease in complexity. In our model, VMs are the buyers who bid for resources at the servers, which are the sellers. From now on, we'll use the term buyers interchangeably with VMs and sellers with servers. We express each buyer as a 4-tuple, $B_i = <\beta_i, \zeta_i, \xi_i, \rho_i>$ where β_i is the bid-per-resource which B_i places, ζ_i is the server which has been allocated to the buyer B_i , ξ_i is the resource requirement of the buyer, and ρ_i is the total payment the buyer will have to pay after the auction is over. We express each seller as a 7-tuple, $S_j = <\alpha_j, \Pi_j, \tau_j, \gamma_j, \rho_j, \epsilon_j^a, \epsilon_j^{id}>$ where α_j is the ask-per-resource which S_j submits, Π_j is the list of VMs to which the given server has been allocated or in other words, the list of VMs hosted at the server, τ_j is the total number of resources at the seller, γ_j is the number of available resources, ρ_j is the total payment the seller will receive after the auction is over, ϵ^a is the

energy the server will consume when it is fully utilized, and ϵ^{id} is the energy the server will consume when it is fully idle. We define utility functions for our buyers and sellers as -

- Buyer Utility Function

$$\mathbb{U}_i^B = \begin{cases} \xi_i v_i - \rho_i, & \text{if } B_i \in \mathbb{B} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where v_i is the true valuation-per-instance for buyer, which actually is the true cost the buyer is willing to pay for a single resource.

- Seller Utility Function

$$\mathbb{U}_j^S = \begin{cases} \rho_j - (\tau_j - \gamma_j)c_j, & \text{if } S_j \in \mathbb{S} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where c_j is the true cost-per-instance of the seller, which actually is the true cost the seller has to incur for lending a single resource to any buyer.

B. Properties of a Double Auction

1) *Individual Rationality (IR)*: Neither any buyer, nor any seller should sustain a loss after joining the auction. This means, for all buyers and sellers winning in the auction,

$$\rho_i \leq \beta_i, \rho_j \geq \alpha_j \quad \forall B_i \in \mathbb{B}, S_j \in \mathbb{S} \quad (3)$$

2) *Balanced Budget (BB)*: Balanced budget deals with the difference between the payments of the buyers and the sellers. It is of 2 types -

- 1) *Strong Budget Balanced (SBB)*: The auctioneer should sustain neither any losses nor gain any profits after the auction. All transactions should be strictly between the buyers and the sellers. The equation for this condition is -

$$\sum_{i=1}^{\eta} \rho_i = \sum_{j=1}^{\kappa} \rho_j \quad (4)$$

- 2) *Weak Budget Balanced (WBB)*: The auctioneer should not sustain any losses after the auction. In other words, the auctioneer should not pay for the transaction to be successful. The auctioneer may however, gain money from the auction. The equation for this condition is -

$$\sum_{i=1}^{\eta} \rho_i \geq \sum_{j=1}^{\kappa} \rho_j \quad (5)$$

3) *Incentive Compatible or Truthful (IC)*: Truthfulness simply means that every buyer's bidding amount should be equal to the true valuation of the resource according to it, and every seller asks for an amount equal to the true cost of its resources. This true bid (resp. ask) submission must be a dominant strategy for all buyers (resp. sellers).

4) *Economic Efficiency (EE)*: After the auction, the seller's products should be in the hands of the buyers who value them the most.

However, according to Myerson–Satterthwaite theorem [3], it is not possible to achieve all four of these simultaneously.

C. Design Objectives

We aim to design a truthful, economic efficient and individually rational double auction scheme to allocate servers to the maximum number of buyers who request resources at a single data center, while minimizing the total energy consumed, once all possible buyers have been allocated and working. Our auction mechanism has two objectives -

- 1) Maximizing the revenue from the VM allocation, while keeping the auction incentive compatible for the buyers.

$$\text{maximize} \sum_{i=1}^{\eta-1} \xi_i q_i + \xi_\eta \beta_\eta \quad (6)$$

where

$$q_i = \begin{cases} \alpha_{\zeta_i}, & \text{if } \alpha_{\zeta_i} \geq \beta_{\delta+1} \\ \beta_{\delta+1}, & \text{otherwise} \end{cases} \quad (7)$$

where δ is the position of B_i in Δ , subject to the following two constraints -

a) *Resource Constraint*

The sum of the resource requirements of all the VMs hosted at a particular server should be less than or equal to the total number of resources available at that server.

$$\forall 1 \leq j \leq \kappa, \sum_{k \in \Pi_j} \xi_k \leq \tau_j \quad (8)$$

b) *Allocation Constraint*

A particular VM can either be allocated to no servers (when it loses), or only one of the servers present at the current data center.

$$\forall B_i \in \mathbb{B}, \zeta_i \in [1, \kappa] \quad (9)$$

We do this by maximizing the size of winning buyers (\mathbb{B}).

$$\text{maximize } |\mathbb{B}| : \mathbb{B} \subseteq \Theta \quad (10)$$

Our basic idea behind maximizing the VM allocation is to allocate expensive seller resources to the buyers with a higher valuation, thus who can afford them. This will leave the cheaper seller resources for the buyers with a lower valuation. This is a slight modification of the Natural Ordering scheme used so often in Double Auction mechanisms.

- 2) Minimizing energy consumption after all possible allocations are done. To do so, we need to select servers such that the resulting combination of the idle and active energies is the least.

$$\text{minimize} \sum_{j=1}^{\kappa} (\epsilon_j^g \psi_j + \epsilon_j^{id} (1 - \psi_j)) \quad (11)$$

where ψ is the server utilization defined as -

$$\psi_j = \frac{\tau_j - \gamma_j}{\tau_j} \quad (12)$$

Theorem III.1. TDAM is NP-hard

Proof. The bin packing problem is: Given n items of different weights $w_1, w_2, w_3, \dots, w_n$ and bins with capacity c , allocate each item to a bin so that the number of total used bins is reduced. The problem statement for our double auction problem is: Given η VMs requiring different number of resources $\xi_1, \xi_2, \xi_3, \dots, \xi_\eta$ and κ servers with capacities $\tau_1, \tau_2, \tau_3, \dots, \tau_\kappa$, assign each VM to a server such that number of assigned VMs is maximized and the energy consumption of the servers is minimized. We can consider this problem with and without server limitations.

- Case 1 : Each server has unlimited resources: Every VM can be placed on the server with the least ask-per-instance energy consumption ratio, say S_j , and we just need to select the buyers with $\beta_i \geq \alpha_j$ and compute the buyers' payment ρ_i as per their bid-per-instance, which can be done in polynomial time.
- Case 2 : Each server has limited resources: We need to place the VMs at the servers according to the Algorithm 1 which is

equivalent to placing η VMs on κ servers such that the number of placed VMs gets maximized, $\sum_{j=1}^{\kappa} e$ gets minimized, the number of used servers gets minimized, and the Resource Constraint and Allocation Constraint are obeyed. Therefore, this problem is equivalent to the Bin-Packing problem. Afterwards, we compute the buyers' and sellers' payment ρ_i and ρ_j according to the Algorithm 2, which is polynomial time solvable.

Here, we can say that our problem can be reduced to the bin-packing problem which is a NP hard. Thus the proposed problem is NP hard too. ■

IV. TDAM BASED VM ALLOCATION FOR REVENUE-ENERGY TRADE-OFF

A. Winner Determination and Allocation

Algorithm 1 Winner_Determination

Input: Buyers set B , Sellers set S

Output: $\Theta, \Gamma, \eta, \kappa$

Phase 1: Initialization

- 1: Build a set $\Theta \leftarrow \phi$ of vertices for B
- 2: Build a set $\Gamma \leftarrow \phi$ of vertices for S
- 3: **for** all buyer B_i in input **do**
- 4: Initialize $\zeta_i \leftarrow -1, \rho_i \leftarrow 0$
- 5: Add $\Theta \leftarrow B_i$
- 6: **for** all seller S_i in input **do**
- 7: Initialize $\Pi_j \leftarrow \text{null}, \gamma_j \leftarrow \tau_j, \rho_j \leftarrow 0$
- 8: Add $\Gamma \leftarrow S_i$
- 9: Set $\eta = \text{number of buyers}$ and $\kappa = \text{number of sellers}$
- 10: All buyers are sorted in Θ to get an ordered list $\Delta = [B_1, B_2, \dots, B_\eta]$ such that $\beta_1 \geq \beta_2 \geq \dots \geq \beta_\eta$
- 11: All sellers are sorted in Γ to get an ordered list $\Omega = [S_1, S_2, \dots, S_\kappa]$ such that $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_\kappa$
- 12: Set $\lambda \leftarrow 0$
- 13: Create 2 sets (unique elements) of winning buyers and sellers resp $\mathbb{B} \leftarrow \phi$ and $\mathbb{S} \leftarrow \phi$

Phase 2: Allocation

- 14: **for** all buyer $B_i \in \Delta$ **do**
- 15: **while** $\lambda \leq \kappa$ and B_i is not allocated **do**
- 16: **if** $\alpha_\lambda \leq \beta_i$ **then**
- 17: **for** all $\mu \in [\lambda, \kappa]$ and until B_i is allocated **do**
- 18: **if** resource requirement of $B_i \leq$ resource availability of S_μ and $\alpha_\mu \leq \beta_i$ **then**
- 19: Set $\zeta_i \leftarrow S_\mu$ ▷ Allocating seller S_μ to buyer B_i
- 20: Add B_i to the client list of S_μ

Phase 3: Updating

- 21: Set $\gamma_\mu \leftarrow \gamma_\mu - \xi_i$ ▷ Decrease available resources at S_μ by ξ_i
- 22: Add B_i and S_μ to \mathbb{B} and \mathbb{S} resp
- 23: Break from the loop
- 24: **else**
- 25: Increment λ

In our double auction mechanism, we first form the lists Θ and Γ from the input and set the values of η and κ

$$\eta = |\Theta|, \kappa = |\Gamma| \quad (13)$$

Then we sort all buyers in non-increasing order of their per-resource bids, in order to satisfy the resource request of the buyer with the highest bid first. We call this new ordered list Δ . Next, we

define a quantity called *Seller Desirability* (σ) which is the ratio of the ask-per-resource of a seller to its *Fully Utilized Energy Consumption Value* (ϵ^{id}), which is the energy consumption when the server is fully utilized.

$$\sigma = \frac{\alpha}{\epsilon^{id}} \quad (14)$$

Thus, we sort all sellers in non-increasing order of their σ -values so as to allocate maximum possible VMs on the server with the least active energy consumption as compared to its idle energy consumption, and the server with the highest ask-per-resource, since the VMs with the highest bids are allocated first. This σ -value is the key to achieving our objective.

Then, we iterate over both the ordered lists to ensure that maximum possible VMs are allocated at a single server, and that too, a server with minimum active energy consumption. We form the list of winning buyers (resp. sellers) and denote it by \mathbb{B} (resp. \mathbb{S}).

B. Pricing Model

Algorithm 2 VCG_based_Payment_Calculation

Input: $\Gamma, \mathbb{B}, \mathbb{S}, \Delta, \Omega$

Output: \mathbb{B}, \mathbb{S}

- 1: All sellers are sorted in Γ to get an ordered list $\Gamma^\alpha = [S_1, S_2, \dots, S_\kappa]$ such that $\alpha_1 \leq \alpha_2 \leq \dots \leq \alpha_\kappa$
 - 2: **for all** seller $S_j \in \mathbb{S}$ **do**
 - 3: **for all** buyer $B_i \in$ client list of S_j **do**
 - 4: **Phase 1: Buyer Payment**
 - 5: **if** B_i is not the last buyer of Δ **then**
 - 6: Set $\rho_i \leftarrow \xi_i * \max(\alpha_j, \beta_{\delta+1})$ where δ is the position of B_i in Δ
 - 7: **else**
 - 8: Set $\rho_i \leftarrow \xi_i * \beta_i$
 - 9: **Phase 2: Seller Payment Revision**
 - 10: **if** S_j is not the last seller of Γ^α **then**
 - 11: Set $\rho_j \leftarrow \rho_j + (\xi_i * \min(\beta_i, \alpha_{\omega+1}))$ where ω is the position of S_j in Γ^α
 - 12: **else**
 - 13: Set $\rho_j \leftarrow \rho_j + \rho_i$
-

Vickrey–Clarke–Groves Mechanism, better known as VCG Mechanism is the most widely used mechanism for payment determination. VCG is not budget balanced, as the auctioneer subsidizes the trade in exchange of a better Individual Rationality. Our payment strategy is inspired from VCG as well. We iterate through the whole list of clients for every server and find the payment for each VM hosted there after allocation by multiplying its resource requirement by the minimum value needed to claim that server, which will be equal to the greater value between the ask of the server and the second highest bid-per-resource. Assuming that δ is the position of the current buyer in the ordered list Δ and the buyer is hosted at server j , we have

$$\rho_i = \begin{cases} \xi_i \alpha_j, & \text{if } \alpha_j \geq \beta_{\delta+1} \\ \xi_i \beta_{\delta+1}, & \text{otherwise} \end{cases} \quad (15)$$

Similarly, we calculate the payment to the seller S_j from its client buyer B_i by finding the minimum value between the buyer's payment and the payment if the given buyer was hosted at the next cheapest server. Assuming that ω is the position of the current seller in the ordered list Γ^α , we have,

$$\rho_j^i = \begin{cases} \xi_i \beta_i, & \text{if } \beta_i \leq \alpha_{\omega+1} \\ \xi_i \alpha_{\omega+1}, & \text{otherwise} \end{cases} \quad (16)$$

The final payment to S_j will equal the sum of payments to S_j from all its clients.

$$\rho_j = \sum_{i=1}^{|\Pi|} \rho_j^i \quad (17)$$

V. THEORETICAL ANALYSIS

A. Energy Approximation

Definition V.1. Asymptotic Approximation Ratio (∇) for any algorithm $A(\Theta, \Gamma)$ is defined as -

$$\nabla_A = \frac{A(\Theta, \Gamma)}{OPT(\Theta, \Gamma)} \quad (18)$$

Total energy consumption is given by -

$$E_{total} = \sum_{j=1}^{\kappa} (\epsilon_j^a \psi_j + \epsilon_j^{id} (1 - \psi_j)) \quad (19)$$

In an ideal scenario, an energy optimal algorithm would not take the asking price of the server into consideration, and would allocate VMs such that minimum energy is consumed. So, the energy consumed in that case would be -

$$E_{OPT} = \sum_{j=1}^{\kappa_a} \epsilon_j^a + \sum_{j=\kappa_a+1}^{\kappa} \epsilon_j^{id} \quad (20)$$

where κ_a is the number of active servers. We assume that the servers are arranged in increasing order of their active energy consumption and decreasing order of their idle energy consumption. So, we can write E_{OPT} as -

$$E_{OPT} = \pi^a + \pi^{id} \quad (21)$$

where π^a is the minimum active energy consumption and π^{id} is the minimum idle energy consumption.

Theorem V.1. The maximum energy consumed by the system of VMs and servers allocated by our mechanism is when $\psi_j = 1, \forall 1 \leq j \leq \kappa$ and is given by

$$\pi^a + \pi^{id} + \sum_{j=\kappa_a+1}^{\kappa} e_j$$

Proof. From Equation 11, we know that

$$E_{TDAM} = \sum_{j=1}^{\kappa} (\epsilon_j^a \psi_j + \epsilon_j^{id} (1 - \psi_j)) \quad (22)$$

Expanding, we get

$$E_{TDAM} = \sum_{j=1}^{\kappa} (\epsilon_j^a \psi_j + \epsilon_j^{id} - \epsilon_j^{id} \psi_j) \quad (23)$$

Here, we define a quantity e , which is the *Energy Consumption Difference Value*, which will always be positive, as it is obvious that a server's active energy consumption will always be greater than its idle energy consumption.

$$e = \epsilon^a - \epsilon^{id} \quad (24)$$

Putting Equation 24 in the above equation,

$$E_{TDAM} = \sum_{j=1}^{\kappa} \psi_j e_j + \sum_{j=1}^{\kappa} \epsilon_j^{id} \quad (25)$$

Now, we assume κ_a to be the number of active servers in the ideal scenario. So obviously, $1 \leq \kappa_a \leq \kappa$. Note that we are not implying that only κ_a servers will be active in the actual scenario in which

our mechanism functions. κ_a is just a variable we are using in our mathematics. Using κ_a in the above equation,

$$E_{TDAM} = \sum_{j=1}^{\kappa} \psi_j e_j + \sum_{j=1}^{\kappa_a} \epsilon_j^{id} + \sum_{j=\kappa_a+1}^{\kappa} \epsilon_j^{id} \quad (26)$$

Introducing a term $\sum_{j=1}^{\kappa_a} \epsilon_j^a$ into the above equation,

$$E_{TDAM} = \sum_{j=1}^{\kappa} \psi_j e_j + \sum_{j=1}^{\kappa_a} \epsilon_j^{id} - \sum_{j=1}^{\kappa_a} \epsilon_j^a + \sum_{j=\kappa_a+1}^{\kappa} \epsilon_j^{id} + \sum_{j=1}^{\kappa_a} \epsilon_j^a \quad (27)$$

Using Equation 24 and Equation 20 in the above equation,

$$E_{TDAM} = \sum_{j=1}^{\kappa} \psi_j e_j - \sum_{j=1}^{\kappa_a} e_j + E_{OPT} \quad (28)$$

Now, since the active and ideal energy values of all the servers are known and constant, the value of E_{TDAM} depends solely on the value of $\psi_j \forall 1 \leq j \leq \kappa$. Since ψ is the server utilization, we know that $\forall 1 \leq j \leq \kappa, 0 \leq \psi_j \leq 1$. Hence, using Equation 21,

$$E_{TDAM} \leq \sum_{j=\kappa_a+1}^{\kappa} e_j + \pi^a + \pi^{id} \quad (29)$$

Corollary V.1.1. *The maximum value of the asymptotic approximation ratio (∇_{TDAM}) of our algorithm is*

$$1 + \frac{\sum_{j=\kappa_a+1}^{\kappa} e_j}{\pi^a + \pi^{id}}$$

Proof. From Equation 18 and Theorem V.1,

$$\nabla_A = \frac{E_{TDAM}}{E_{OPT}} \leq 1 + \frac{\sum_{j=\kappa_a+1}^{\kappa} e_j}{\pi^a + \pi^{id}} \quad (30)$$

B. Time Complexity Analysis

Theorem V.2. *The worst case time complexity of TDAM is $\mathcal{O}(5\eta\kappa + 2\eta \log \eta + \eta \log \kappa)$*

Proof. The complexity of TDAM will be the sum of complexities of Winner determination and Payment calculation algorithm. Assuming that $\eta \gg \kappa$, we have the worst case time complexity of TDAM to be $\mathcal{O}[(2\eta + 2\kappa) + (\eta \log \eta + \kappa \log \kappa + 5\eta\kappa + 2\eta) + (\kappa \log \kappa + \eta \log \eta + \eta \log \kappa)] \approx \mathcal{O}[5\eta\kappa + 2\eta \log \eta + \eta \log \kappa]$

C. Properties of TDAM

Theorem V.3. *TDAM is individually rational.*

Proof. An auction mechanism will be individually rational if the buyer pays less than or equal to its bid, and the seller gets greater than or equal to its ask. We know that $\forall B_i \in \Theta$, there are 2 possibilities

- 1) B_i is not a winning buyer: We need not discuss its rationality.
- 2) B_i is a winning buyer: We have 2 more cases according to Algorithm 2 -

- a) B_i is not the last buyer of Δ .

We know that $\alpha_j \leq \beta_i$, because that is the criteria for B_i to be a winning bid. Also, we know that $\beta_{\delta+1} \leq \beta_i$ since the buyers in Δ are arranged in decreasing order of their bid-per-instance. So, $\rho_i \leq \xi_i \beta_i$.

- b) B_i is the last buyer of Δ .

$$\rho_i = \xi_i \beta_i.$$

We know that $\forall S_j$ in Γ , there are 2 possibilities -

- 1) S_j is not a winning seller: We need not discuss its rationality.

- 2) S_j is a winning seller: We have 2 more cases according to Algorithm 2 -

- a) S_j is not the last seller of Γ^α .

We know that $\beta_i \geq \alpha_j$, because that is the criteria for B_i to be a winning bid. Also, we know that $\alpha_{\omega+1} \geq \alpha_j$ since the sellers in Ω are arranged in increasing order of their ask-per-instance. So, $\rho_j \geq (\tau_j - \gamma_j)\alpha_j$.

- b) S_j is the last buyer of Ω .

$$\rho_j = (\tau_j - \gamma_j)\alpha_j.$$

Lemma V.4. *TDAM is truthful for buyers and sellers*

Proof. This proof is omitted due to limited space. ■

Theorem V.5. *TDAM is economic efficient.*

Proof. Since we assign a higher priority to a buyer with a higher bid-per-instance, the seller's resources go to the buyers which value them the most. So, TDAM is economic efficient. ■

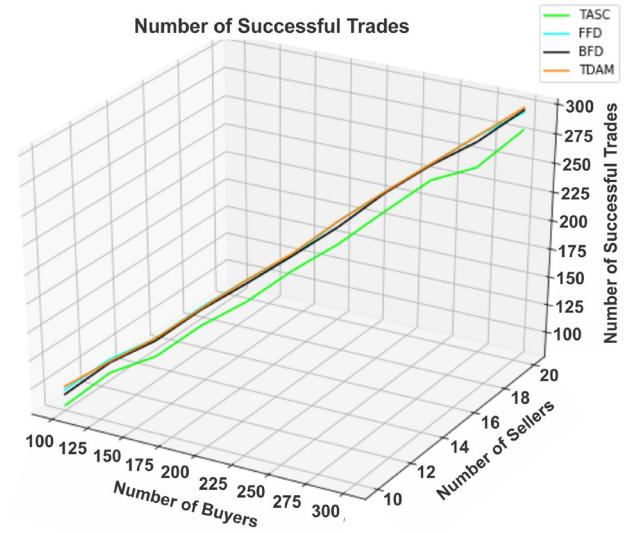


Figure 2. Number of Successful Trades

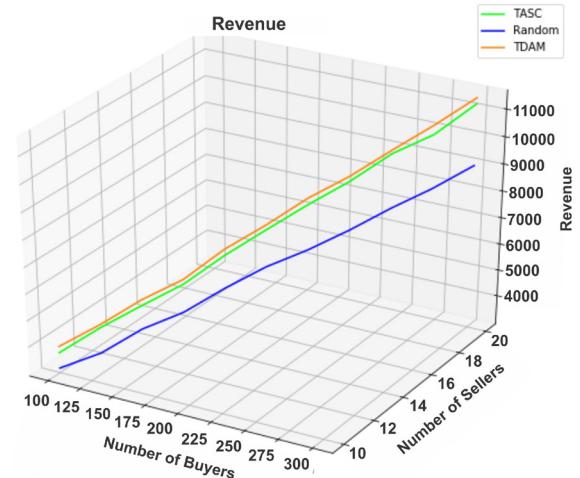


Figure 3. Revenue for fixed number of buyers and sellers

VI. SIMULATION ANALYSIS

To evaluate performance of the proposed mechanism. We have randomly generated uniformly distributed values for the variables in the following ranges: $\beta \in [2.0, 12.0]$, $\xi \in [1, 11]$, $\alpha \in [1.0, 10.0]$, $\tau \in [\eta, 2\eta]$, $\epsilon^{id} \in [150.0, 250.0]$, and $\epsilon^a \in [\epsilon^{id} + 10.0, 300.0]$. We compare proposed algorithm with four other well-known allocation algorithms: TASC [12], First-Fit Decreasing (FFD), Best-Fit Decreasing (BFD), and Random Allocation. We perform these comparisons based on the following: 1) *Number of Successful Trades*: As shown in figure 2, TDAM works way better than TASC and slightly better than FFD and BFD in giving the maximum successful trades when number of VMs and PMs are fixed. Since Random Allocation will eventually consider all VMs and pair them with PMs randomly, it allocates more VMs than our algorithm when time is not considered. 2) *Revenue Analysis*: As shown in figure 3, the revenue generated from VM allocation by TDAM for fixed number of buyers and sellers is the maximum among the three revenue generating algorithms we took into consideration. 3) *Energy Consumption Analysis*: As depicted in figure 4, the energy consumption of the data center, which is equal to the total energy consumption of all servers is minimum for maximum successful trades when TDAM is used for allocation. 4) *Running Time*: As shown in figure 5, all the algorithms perform almost similar for small number of VMs and PMs, but when number of VMs is increased over 10,000 and PMs is increased over 100, TDAM starts to yield results faster than other algorithms.

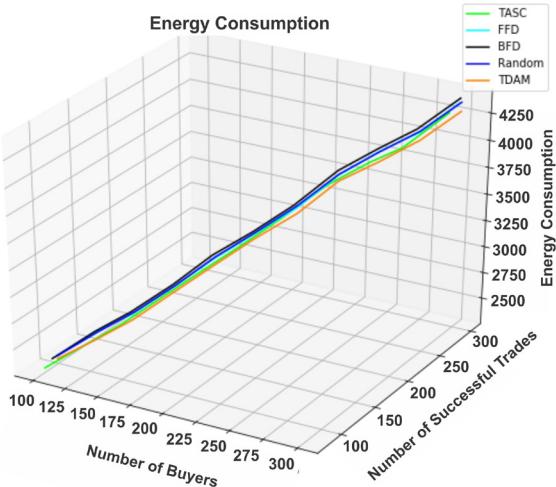


Figure 4. Energy Consumption while maximizing successful trades

VII. CONCLUSIONS

We have designed TDAM, a truthful double auction mechanism based resource allocation for revenue-energy trade-off in cloud data centers. In the proposed auction model VMs work as the buyers and servers work as the sellers of resources. To ensure the truthfulness, TDAM model explicitly enforce both buyers and sellers to submit their true valuations and also satisfy other economical properties such as economic efficiency and individual rationality. We have proposed two algorithms one is for winning bids determination which is converted into minimum weighted bipartite matching and other is a VCG-based auction for payment determination of winning bids. Experimental results validate our theoretical analysis and show that in comparison with existing auction approaches TDAM can significantly

improve the performance in terms of revenue, number of active servers, energy consumption and execution time.

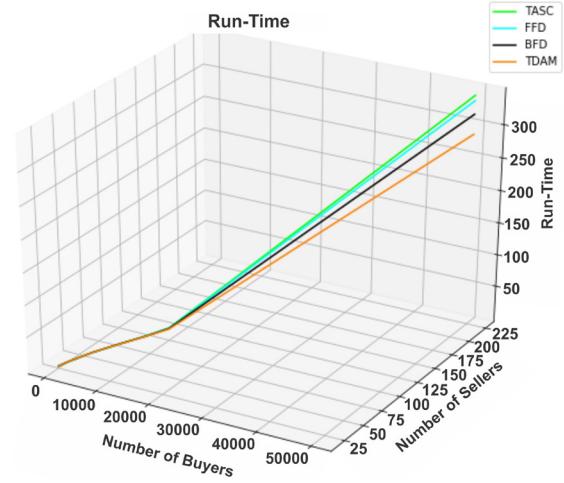


Figure 5. Run-Time trend as numbers of VMs and PMs are increased

REFERENCES

- [1] Michael Armbrust et al., "A view of cloud computing", Commun. ACM 53, 4, April 2010, pp. 50-58.
- [2] V. Krishna,"Auction Theory", Academic Press, March, 3,2002.
- [3] Roger B Myerson, Mark A Satterthwaite, "Efficient mechanisms for bilateral trading", Journal of Economic Theory, vol. 29, no.2, 1983, pp. 265-281,
- [4] X. Wang, X. Chen and W. Wu, "Towards truthful auction mechanisms for task assignment in mobile device clouds," IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, Atlanta, GA, 2017, pp. 1-9.
- [5] X. Wang, X. Chen, W. Wu, N. An and L. Wang, "Cooperative Application Execution in Mobile Cloud Computing: A Stackelberg Game Approach," in IEEE Communications Letters, vol. 20, no. 5, 2016, pp. 946-949
- [6] L. Jin, W. Song, P. Wang, D. Niyato and P. Ju, "Auction Mechanisms Toward Efficient Resource Sharing for Cloudlets in Mobile Cloud Computing," in IEEE Transactions on Services Computing, vol. 9, no. 6, 2016,pp. 895-909.
- [7] L. Mashayekhy, M. M. Nejad, D. Grosu and A. V. Vasilakos, "Incentive-Compatible Online Mechanisms for Resource Provisioning and Allocation in Clouds," 2014 IEEE 7th International Conference on Cloud Computing, Anchorage, AK, 2014, pp. 312-319.
- [8] W. Wang, Y. Jiang and W. Wu, "Multiagent-Based Resource Allocation for Energy Minimization in Cloud Computing Systems," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 47, no. 2, 2017, pp. 205-220.
- [9] Debasis Mishra, Rahul Garg, "Descending price multi-item auctions", Journal of Mathematical Economics, vol. 42, Issue 2, 2006, pp. 161-179.
- [10] Demange, Gabrielle and Gale, David and Sotomayor, Marilda,"Multi-Item Auctions",Journal of Political Economy,vol. 94,no. 4,1986, pp. 863-872.
- [11] M. Ausubel et al., "An efficient dynamic auction for heterogeneous commodities",The American Economic Review,2006,pp. 602-629.
- [12] Dejun Yang, Xi Fang, and Guoliang Xue, "Truthful auction for cooperative communications", In Proceedings of the Twelfth ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '11). ACM, New York, NY, USA,Article 9, 2011.
- [13] Huang, P. , Scheller-Wolf, A. and Sycara, K., "Design of a Multi-Unit Double Auction E-Market", Computational Intelligence, 18, 2002, pp. 596-617
- [14] Charles R. Plott, Peter Gray, "The multiple unit double auction", Journal of Economic Behavior & Organization, vol. 13, Issue 2, 1990, pp. 245-258

Development of Assamese Text-to-speech System using Deep Neural Network

Abhash Deka*, Priyankoo Sarmah*, Samudravijaya K*, S R M Prasanna*†

*Center for Linguistic Science and Technology

Indian Institute of Technology Guwahati, Guwahati, Assam 781039 INDIA

Email: {abhash, priyankoo, samudravijaya}@iitg.ac.in

†Department of Electronics and Electrical Engineering

Indian Institute of Technology Dharwad, Dharwad, Karnataka 580011 INDIA

Email: prasanna@iitdh.ac.in

Abstract—This paper describes the development of a text-to-speech system for Assamese language, using Deep Neural Network (DNN). The system is trained with speech data, collected by a consortium, that is available free of cost for academic use. The DNN based method eliminates the need for a grapheme to phoneme conversion; rather, it synthesizes speech directly from the UTF-8 based Assamese script. The results of objective and subjective evaluations confirm that the Assamese speech synthesized using DNN approach is better than the ones synthesized using the traditional hidden Markov model based text-to-speech system.

I. INTRODUCTION

Development of text-to-speech (TTS) systems in Indian languages received a fillip from the Indian languages TTS consortium project sponsored by MeitY, Government of India [1]. The efforts of the TTS consortium resulted in the development of TTS systems in 13 Indian languages, namely, Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Gujarati, Odia, Assamese, Manipuri, Kannada, Bodo and Rajasthani [2].

Initial versions of the TTS systems followed concatenative or unit selection synthesis approaches, which use small units of speech sounds to produce synthetic speech [2]. However, to produce all speech sounds in a variety of acoustic-phonetic and prosodic contexts, a large speech database containing all possible phonemes and their combinations is required to be stored within the machine. This renders the task difficult for a low-resource language such as Assamese. However, by using statistical parametric speech synthesis methods, such as the Hidden Markov Model (HMM), the need to store large amount of speech sound files is obviated. Rather, this approach stores parameters, such as fundamental frequency and cepstral coefficients, of statistical models of speech sounds in order to synthesize speech corresponding to the input text. This method is superior to the concatenation based approach [3] as it has low requirement of training data, useful for low resource languages. Moreover, this approach has the ability to change voice characteristics, such as the speaking style of the speaker and speaker emotions. However, in this method, the quality of synthesized speech is not as natural as that of the concatenative approaches, requiring additional processing of the synthesized speech. Zen et al. described three reasons that are responsible

for degraded speech: vocoding, accuracy of acoustic models and over-smoothing [4].

To reduce such degradation in speech quality, and to improve the naturalness of synthesized speech, Deep Neural Network (DNN) based approaches have been used in recent years. Availability of large datasets and improvement in computational power has made the use of DNN for TTS easier. Also, recent studies show that building TTS systems using deep neural network results in more natural synthesized speech [5]. This is due to the ability of the DNN to model long-span frames [6], and to model high dimensional and strongly correlated features, and to find a highly non linear mapping between input and output features. An investigation revealed that primary reasons for improved naturalness ratings of humans listening to synthetic speech produced by TTS systems using DNNs were the replacement of decision trees with DNNs and moving from state-level to frame-level predictions [7].

Most of the 13 language TTS systems, implemented by TTS consortium, were initially developed using the festvox framework [2], [8]. Hidden Markov model based text-to-speech (HTS) engines were developed subsequently [3]. The development of HTS systems was facilitated by TBT, an open source toolkit to build multiple language TTS systems [9]. A by-product of the TTS consortium project was the creation of a speech database in many Indian languages for implementing TTS systems [1]. We used the Assamese language module of this speech database to build a new TTS system using DNN to model the mapping between text based features and the corresponding acoustic features needed for generation of speech signal. The current system generates speech waveform corresponding to an input Assamese text in UTF-8 format. The quality of the speech generated by this DNN based system is better than those generated by TTS systems following the conventional statistical parametric approach using hidden Markov model. Here, we report the development of a DNN based Assamese TTS system, and subjective as well as objective evaluation of the quality of speech synthesized by the system.

The rest of the paper is organized as follows. The speech database and the toolkit used for implementing the TTS system are described in Section II. The details of the subjective and objective evaluation methodologies are also given in the same

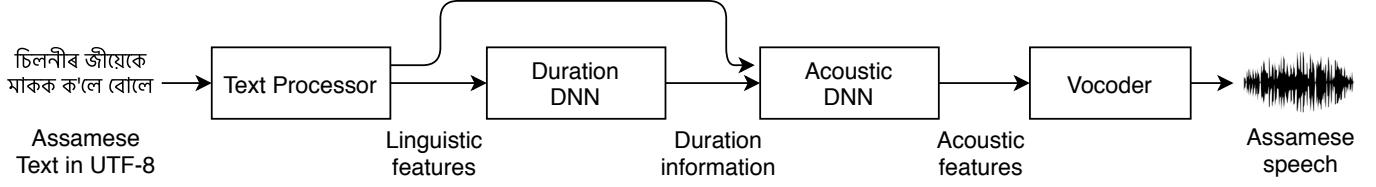


Fig. 1. Block diagram of the Assamese TTS system

section. The results of the evaluation process are presented and discussed in section III. Concluding remarks are drawn in section IV.

II. DATABASE AND METHODOLOGY

In this section, we describe the speech data and the transcriptions used to train the DNN based Assamese TTS system. The details of the implementation of the TTS system, and methodology of conducting objective as well as subjective evaluation tests are also given in this section.

A. Database

To build our text-to-speech synthesis system for Assamese language, we used the speech database of the TTS consortium. The speech data can be downloaded for academic and research use [10]. The Assamese database contains sound files generated by a male and a female speaker. We used 8941 audio files, spoken by a male speaker, to build the TTS system. Each sound file contains one spoken sentence. The digitized speech is stored in raw format (48000 Hz, signed 16-bit PCM, little endian, mono). The total duration of recorded speech data is 12.95 hours. A single text file contains the text corpus comprising of 8,941 sentences. We created 8,941 text files, each containing the sentence corresponding to one sound file. The text corpus contains 32,136 unique words [10]. The speech files and the corresponding transcriptions were used to build the Assamese text-to-speech system using DNNs as described below.

B. Implementation of TTS system based on DNN

Figure 1 shows a block diagram of the Assamese TTS system. It contains 4 processing blocks: (i) a text processor to generate linguistic specification of the input Assamese text, (ii) a (duration) DNN to generate duration information of the textual unit from its linguistic specification, (iii) another (acoustic) DNN to map the linguistic features to acoustic features, and (iv) a vocoder to generate speech data from the acoustic features and duration information of the textual unit. A brief description of the task carried out by each processing block as well as the software tools used to carry out the tasks is provided below. For details of application of DNNs for TTS, one may read a tutorial presented at Interspeech 2017 [11].

1) Text Analysis: In order to derive linguistic features from Assamese text in UTF-8 format, we have used Ossian [12] as a front end. Ossian is a collection of python codes that aids in building TTS systems. It is an open source toolkit, distributed using Apache License 2.0. Ossian supports the

use of neural nets trained with the Merlin toolkit [13] as duration and acoustic models. Ossian relies on the HMM toolkit (HTK) [14] and HMM based text-to-speech (HTS) system [15] for alignment of speech data with transcription. The biggest advantage of Ossian is that it does not require any language specific knowledge to extract the linguistic features. The UTF-8 characters are used to tokenise the text, and characterise tokens as words, white space, punctuation etc. Ossian uses a letter/grapheme based approach in which the names of letters are used directly as the names of speech modelling units in contrast to the traditional approach where phonemes are used as the speech units. This approach eliminates the need to have any language specific knowledge such as phonetic categories (vowel, nasal, approximant etc.) as well as part of speech categories (noun, verb, adjective etc.).

Using Ossian, the following speech features were extracted from the Assamese speech data: mel-generalized cepstrum, logarithm of fundamental frequency and mean band aperiodicity. The output is formatted as HTS-style labels using state-level alignment. The labels are then converted into sequence of vectors of binary and continuous features by employing HTS-style questions to derive the features from the label sequence. The resulting linguistic features are used as input features for training duration and acoustic DNNs.

2) Deep Neural Networks: The current TTS system takes linguistic features as input, and employs a combination of two DNNs to predict acoustic features, which are then passed to a vocoder to produce the speech waveform. One (duration) DNN was trained to predict the duration of letter units from acoustic features. Another (acoustic) DNN with 6 hidden layers was trained to map the input linguistic features and associated duration features into acoustic features.

The sequence of vectors of binary and numeric features generated by the text processor were normalized in the range of [0.001, 0.99] before feeding to the input layer of DNN. The duration DNN was trained using the (forced Viterbi) time-aligned data, to predict duration of a state of a HMM or the duration of an entire HMM representing a quin-letter (a letter with 2 left and 2 right letter contexts).

The acoustic features at the output layer of acoustic DNN are Line Spectral Pairs (LSP), Fundamental frequency (F0) and unvoiced/voiced (U/V). The DNN was trained using Stochastic Gradient Descent algorithm, with an initial value of the learning rate as 0.02. The weights of DNN were estimated by using frame-aligned pairs of input and output features, extracted from the training data, to minimize errors between

outputs mapped by the DNN and the target outputs [6]. The sequence of output features were normalized to have zero mean and unit variance.

During synthesis, duration of a quin-letter is predicted first. Then, the linguistic features of the quin-letter and its predicted duration are fed to the acoustic DNN to predict the sequence of acoustic feature vectors. Thanks to the duration prediction, typically many acoustic feature vectors get generated by the acoustic DNN corresponding to one input linguistic feature vector.

3) Vocoder: To synthesize the time waveform using the acoustic features estimated by the DNN, we have used WORLD [16], a free vocoder. Merlin toolkit contains a version of WORLD, modified to satisfy the requirements of Merlin.

By setting the predicted output features from the DNN as mean vectors, and using the pre-computed (global) variances of output features from all training data, the speech feature generation module generates smooth trajectories of speech parameter features which satisfy the statistics of static and dynamic features. In WORLD, the vocal cord vibration is calculated on the basis of the convolution of the excitation signal with the minimum phase response of the spectral envelope, interpolated at excitation pulse locations along the time axis [16]. The F0 information is used to determine the temporal positions of the pulse location (origin of each vocal cord vibration).

C. Objective and Subjective Evaluation

In order to assess the expected improvement in quality of speech synthesized by the current system that uses DNN instead of GMM-HMM [8], we carried out both objective and subjective evaluations. For objective evaluation, we adopted the Perceptual Evaluation of Speech Quality (PESQ) [17], [18] as a measure of quality of synthetic speech. For objective evaluation, 100 sentences were randomly chosen from the database of 8941 Assamese sentences. The original Assamese sentences, spoken by a human subject, were the reference items; the corresponding synthesized sentences, using GMM-HMM [8] and DNN approaches, constituted the test set.

Subjective evaluation of the synthesized speech was conducted by 26 human raters, who are native speakers of Assamese. The Assamese speakers performed a Differential Mean Opinion Score (DMOS) task where they were asked to score the quality of the speech generated by a TTS system with reference to the speech of the same text spoken by an Assamese speaker [19]. In this Degradation Category Rating method, a subject listens to the utterance of a sentence as produced by a human followed by the speech of the same sentence synthesized by a TTS system, and rates the relative score of the synthetic speech. In the current work, audio files listened to by human raters, contained natural speech, a 440Hz beep and the corresponding synthetic speech in that order. The sampling frequency of the mono channel audio file was 48kHz.

Of the 100 sentences chosen for the objective analysis, 15 sentences were selected for the subjective evaluation. Speech

data corresponding to each of these sentences was synthesized by both GMM-HMM and DNN based TTS systems. This yields 30 synthetic stimuli. In the evaluation task, a synthetic stimulus was played after playing the corresponding reference/natural speech. Each stimulus was presented twice to each and every subject, resulting in 60 total stimuli, per subject. The 60 stimuli were presented in random order for evaluation.

Each subject was instructed to rate the degradation of the synthesized speech in comparison to the human speech on a scale from 5 to 1, corresponding to the following judgments respectively: imperceptible, perceptible but not annoying, slightly annoying, annoying and very annoying. The methodology followed was in accordance with the guidelines provided in ITU-T Rec. P.913 [20] and discussed in detail by Pinson and Janowski [21].

The perceptual evaluation test was conducted via a Praat [22] based graphical user interface on a laptop computer. The reference and the test speech were normalized to have equal intensity. The subjects listened to the stimuli using a headphone with a flat response in the frequency range 20–20000Hz. The subjects were allowed to listen to the sentences as many times as they liked, before making a judgment. The graphical user interface allowed the subjects to click one of the five buttons to choose the degradation category they deemed appropriate. Each subject took about 20 minutes to complete the DMOS test. The evaluation scores were extracted and tabulated for statistical analysis as detailed in Section III-B.

III. RESULTS AND DISCUSSION

The results of objective and subjective evaluation of the quality of speech generated by the traditional GMM-HMM based TTS system and the current, DNN-based TTS system are presented and discussed in this section.

A. Objective evaluation

The standard objective measure of speech quality, PESQ [18], provides two scores: PESQ-MOS and PESQ-LQO (Mean Opinion Score; Listening Quality Objective). A mapping function maps the PESQ-MOS raw values in the range [-0.5, 4.5] to PESQ-LQO in the range [1, 5]. These two scores for speech synthesized by the two (GMM-HMM and DNN) TTS systems, are tabulated in Table I.

TABLE I
PESQ SCORES OF SPEECH SYNTHESIZED BY TTS SYSTEMS BASED ON
GMM-HMM AND DNN

	MOS	LQO
GMM-HMM	0.3	1.1
DNN	0.7	1.1

Since the difference between the average PESQ scores are marginal, we decided to conduct two separate paired t-tests for PESQ-MOS and PESQ-LQO values. There was a statistically significant difference in the PESQ-MOS scores of the synthesized speech using GMM-HMM and DNN approaches, with $t(99)=8.4$, $p < 0.0001$. Similarly, the two approaches

differed significantly with respect to PESQ-LQO scores; the corresponding $t(99)=5.7$, $p< 0.0001$.

The PESQ-MOS scores of 100 synthetic speech signals are shown in the form of a bar chart in Fig. 2. One can see that the scores of the speech generated by the DNN system (blue bars) are, in general, higher than those of the GMM-HMM system (red bars). Similar trend is visible in the case of PESQ-LQO scores as shown in Fig. 3. In the case of both scores, 83 out of 100 synthesized sentences using the DNN approach have received higher PESQ-MOS and PESQ-LQO scores than the ones synthesized using the GMM-HMM approach.

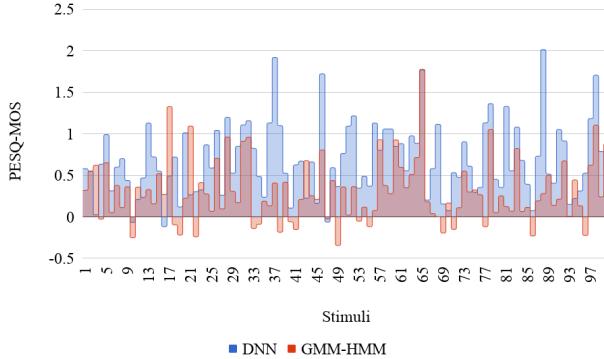


Fig. 2. PESQ-MOS scores of 100 speech signals generated by GMM-HMM (red bar) and DNN (blue bar) based TTS systems.

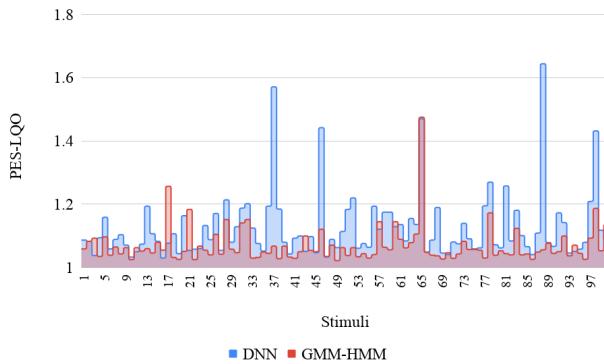


Fig. 3. PESQ-LQO scores for 100 sentences

B. Subjective evaluation

A subjective evaluation of the synthesized stimuli, using DMOS scores as described in Section II-C, showed that all 26 speakers consistently rated the synthesized stimuli generated by the DNN approach as better. The average of DMOS scores corresponding to the GMM-HMM and DNN TTS systems are provided in Table II. The average DMOS score of DNN based TTS system is 1.0 higher than that of the conventional GMM-HMM based TTS system.

Each of the 15 synthetic speech was scored by 26 native Assamese speakers. The average of DMOS scores of all 26

TABLE II
AVERAGE OF DMOS SCORES OF THE TWO TTS SYSTEMS

	DMOS
GMM-HMM	2.7
DNN	3.7

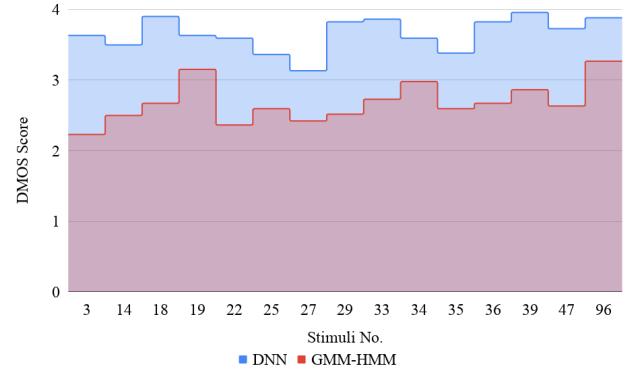


Fig. 4. Subjective evaluation (DMOS) of 15 stimuli synthesized by the GMM-HMM and DNN based TTS systems.

listeners for each of the 15 stimuli generated by GMM-HMM and DNN methods are presented in a stepped chart in Fig. 4. As seen in the figure, in the case of all 15 stimuli, human listeners have rated the quality of the DNN generated stimuli better than the GMM-HMM generated stimuli. A paired t-test confirmed that the DMOS scores obtained by the DNN generated stimuli are significantly different from the ones generated by the GMM-HMM approach [$t(779) = 24.7$, $p< 0.0001$].

IV. CONCLUSION

In this paper, we reported the development of an Assamese text to speech system that uses a deep neural network to map linguistic features of input text to acoustic features that was used by a vocoder to generate good quality speech signal. The Assamese TTS system was trained using the Assamese language module of the speech database of the TTS consortium. Both subjective and objective tests reveal that the quality of the speech synthesized by the DNN based TTS system developed by us is distinctively better than the quality of speech synthesized by the GMM-HMM based TTS system. Our system also obviates the need for grapheme to phoneme conversion, and generates speech corresponding to the text written in Assamese script in UTF-8 format.

ACKNOWLEDGMENT

This work is part of a project titled “Development of Speech Interface for Form-filling application (SiFA) in Five Indian languages”, funded by the Ministry of Electronics and Information Technology (MeitY) and the Ministry of Human Resource Development (HRD), Government of India, under the IMPRINT research grant program. The implementation of the Assamese TTS system was benefited by the Assamese speech data recorded as part of the TTS consortium, sponsored

by MeitY. The authors would like to express their gratitude to Mr. Giridhar Pamishetty of IIT Hyderabad for his help. The authors would also like to express their sincere gratitude towards the reviewers of NCC 2019 for their constructive suggestions.

REFERENCES

- [1] TDIL_TTS, "Indian Language Technology Proliferation and Deployment Centre: Text-to-speech," available: <http://tdil-dc.in/>.
- [2] H. A. Patil, T. B. Patel, N. J. Shah, H. B. Sailor, R. Krishnan, G. R. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S. P. Kishore, S. R. M. Prasanna, N. Adiga, S. R. Singh, K. Anand, P. Kumar, B. C. Singh, S. L. B. Kumar, T. G. Bhadran, T. Sajini, A. Saha, T. Basu, K. S. Rao, N. P. Narendra, A. K. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, and H. A. Murthy, "A syllable-based framework for unit selection synthesis in 13 Indian languages," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Nov 2013, pp. 1–8.
- [3] B. Ramani, S. L. Christina, G. A. Rachel, V. S. Solomi, M. K. Nandwana, A. Prakash, S. A. Shanmugam, R. Krishnan, S. K. Prahalad, K. Samudravijaya, P. Vijayalakshmi, T. Nagarajan, and H. A. Murthy, "A common attribute based unified HTS framework for speech synthesis in Indian languages," in *Proc. of SSW8: 8th ISCA Speech Synthesis Workshop, Barcelona, Spain*. ISCA, 2013, pp. 291–296.
- [4] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*. IEEE, 2013, pp. 7962–7966.
- [5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *CoRR*, vol. abs/1609.03499, 2016, available: <http://arxiv.org/abs/1609.03499>.
- [6] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 3829–3833.
- [7] O. Watts, G. Henter, T. Merritt, Z. Wu, and S. King, "From HMMs to DNNs: Where do the improvements come from?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. IEEE, 2016, pp. 5505–5509.
- [8] B. Sharma, N. Adiga, and S. R. M. Prasanna, "Development of Assamese Text-to-speech synthesis system," in *TENCON 2015 - 2015 IEEE Region 10 Conference*, 2015, pp. 1–6.
- [9] A. S. Ghone, R. Nerpagar, P. Kumar, A. Baby, A. Shanmugam, S. M., and H. A. Murthy, "TBT (Toolkit to Build TTS): A High Performance Framework to Build Multiple Language HTS Voice," in *Proc. Interspeech 2017*, 2017, pp. 3427–3428.
- [10] Arun Baby and Anju Leela Thomas and Nishanthi N L and TTS Consortium, "Resources for Indian languages," in *Proc. CBBLR Community Based Building of Language Resources, Brno, Czech Republic*, 2016, pp. 37–43, available: <https://www.iitm.ac.in/donlab/tts/>.
- [11] S. King, O. Watts, S. Ronanki, W. Zhizheng, and F. Espic, "Deep learning for text-to-speech synthesis, using the merlin toolkit," August 2017.
- [12] "Ossian, a toolkit for building TTS systems," <http://jrmeyer.github.io/tts/2017/09/15/Ossian-Merlin-demo.html>.
- [13] Z. Wu, O. Watts, and S. King, "Merlin: An Open Source Neural Network Speech Synthesis System," *Proc. 9th ISCA Speech Synthesis Workshop (SSW9), September 2016, Sunnyvale, CA, USA*, 2016.
- [14] "HTK, Speech Recognition Toolkit." [Online]. Available: <http://htk.eng.cam.ac.uk/>
- [15] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system HTS version 2.0," in *Proc. of the 6th ISCA Workshop on Speech Synthesis SSW, Bonn, Germany*, 2007, pp. 294–299, available: <http://hts.sp.nitech.ac.jp/>.
- [16] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [17] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, May 2001, pp. 749–752 vol.2.
- [18] I.-T. R. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," International Telecommunication Union, Tech. Rep., 2016, available: <https://www.itu.int/rec/T-REC-P.862>.
- [19] I.-T. R. P.800.2, "Mean opinion score interpretation and reporting," International Telecommunication Union, Tech. Rep., 2016, available: <https://www.itu.int/rec/T-REC-P.800.2/en>.
- [20] ITU, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," International Telecommunication Union, Tech. Rep., 2016.
- [21] M. H. Pinson and L. Janowski, "A new subjective audiovisual & video quality testing recommendation," *VQEG eLetter*, vol. 1, pp. 51–60, 2014.
- [22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. Version 6.0.43, retrieved on 8 september 2018 from <http://www.praat.org/>," 2018.

Multimodal Fusion of Speech and Text using Semi-supervised LDA for Indexing Lecture Videos

Moula Husain

Computer Science and Engineering

B.V.B College of Engineering and Technology

Hubballi, India

moulahusain786@gmail.com

Meena S. M

School of Computer Science and Engineering

K.L.E Technological University

Hubballi, India

msm@kletech.ac.in

Abstract—Lecture videos are the most popular learning materials due to their pedagogical benefits. However, accessing a topic or subtopic of interest requires manual examination of each frame of the video and it is more tedious when the volume and length of videos increases. The main problem thus becomes the efficient automatic segmentation and indexing of lecture videos that enables faster retrieval of specific and relevant content. In this paper, we present automatic indexing of lecture videos using topic hierarchies extracted from slide text and audio transcripts. Indexing videos based on slide text information is more accurate due to higher character recognition rates but, text content is very abstract and subjective. In contrast to slide text, audio transcripts provide comprehensive details about the topics, however retrieval results are imprecise due to higher WER. In order to address this problem, we propose a novel idea of fusing complementary strengths of slide text and audio transcript information using semi-supervised LDA algorithm. Further, we strive to improve learning of the model by utilizing words recognized from video slides as seed words and train the model to learn the distribution of video transcriptions around these seed words. We test the performance of proposed multimodal indexing scheme on 500 number of class room videos downloaded from Coursera, NPTEL and KLETU (KLE Technological University) classroom videos. The proposed multimodal fusion based scheme achieves an average percentage improvement of 44.49% F-Score compared with indexing using unimodal approaches.

Index Terms—Multimodal fusion, Indexing, semi-supervised LDA

I. INTRODUCTION

Lecture videos are the most popular learning materials in e-learning platforms. Due to rapid growth in the video recording systems, compression techniques, cloud and internet technology, most of the universities offer e-learning by recording their class room lecture videos and publishing them in their web portals. As a result, the volume of lecture video content is increasing at exponential rate. Since lecture videos have longer duration, searching a video and accessing a particular topic of interest is time and labor intensive task. It requires manual examination of individual video frames or audio segments to search specific and relevant content. The main problem thus becomes the efficient automatic segmentation and indexing of lecture videos that enables faster retrieval of specific and relevant content.

The last decade witnessed significant efforts in the area of video indexing and retrieval systems. Several attempts were started with segmenting videos manually and annotating each segment with relevant metadata information [20] [3]. The manual annotation and indexing of videos is a simple way to provide access to required content of videos. However, manual indexing is time intensive and has scalability issues for real time applications. Researchers [23] [17] have proposed automatic ways of indexing lecture videos by using single and multiple modalities (text, speech, visual features etc..) of information. Lecture slide text and audio content are the two prominent modalities which convey maximum information and are useful for automatic annotation, structuring and indexing of lecture videos. Extraction of text from lecture video frames involves preprocessing operations such as detection, localization, enhancement, and text recognition. Text recognition from video frames is a well studied problem in the area of pattern recognition using machine learning (KNN, SVM) and deep learning (LSTM, CNN) approaches [13] [9]. Wonjun Kim et al. [12] proposed extraction of overlay text for complex background videos by using transition maps obtained between image text and its background. Tayfun T et al. [21] proposed an Indexed, Captioned and Searchable (ICS) video framework. The ICS framework segments and indexes videos by applying OCR on video frames which are enhanced using image transformations. Text recognition applied to detect titles and subtitles for lecture videos provides an useful information during retrieval operations. However, lecture slides have limited content and can provide mere overview of the concepts. Besides video text, audio contains detailed description of the content and is correlated with lecture slides. In addition to video text, recognition of spoken words from audio plays a crucial role in content based indexing of lecture videos. The state of the art speech recognition tools include Kaldi, CMU Sphinx, Dragon Natural Speaking (DNS) and Julius. The major problem with these ASR engines is high Word Error Rate (WER) in transcribing the lecture audio into text descriptions due to unconstrained audio recording, variation in ascent and large Out-of-Vocabulary (OOV) words. Several applications such as GAudi (Google Audio Indexing) [4], PodCastle [18] and SpeechFind [2] are based on spoken word detection and retrieval. Similarly MIT Lecture Browser [6] and

NTU Virtual Instructor [24] are built based on spoken words. These approaches use lattice based techniques to improve WER of ASR by representing detected spoken words in terms of lattices. Sub-word based techniques were adopted to overcome the problem of OOV words. These applications have shown promising results of 15-20% WER which compares favorably with accuracy of human transcriptions. However, ASR performance is not adequate when it is applied on more challenging real world datasets involving continuous long sequence of audio and OOV words. Several research attempts have been made to improve WER of ASRS [10] [25]. However, achieving low WER for audio content supporting diversity in language, increasing vocabulary, varying ascent and unconstrained audio recording environments still remains as a challenging issue.

It is evident from aforementioned discussion that recognition rate for video text is relatively high but has limited content. In contrast, lecture audio content has low recognition accuracy but contains detailed information about the topics and subtopics. Therefore, indexing videos by combining these complementary and correlative features has the potential to improve retrieval performance. In [7], Matthew Cooper attempted to use visual and aural channels extracting from more than 60 hours of lecture videos. The author has compared documents retrieved based on text information with speech content and proved, indexing videos with slide text provides better precision compared to spoken text. In [11], authors have presented text/captions based segmentation and indexing of lecture videos. Authors have validated segmentation and indexing of lecture videos based on text generated using corrected and uncorrected OCR text. They reported improvement of 11% using corrected ASR text compared to OCR text. In [16], authors have proposed multimodal retrieval (where users are interested to access videos based on the query containing text, graphics or both) and cross modal retrieval (where users want to access graphics by giving text as input or vice versa). Authors have utilized visual text, graphics and speech modalities for segmenting and indexing of videos. In recent work [23], Haojin Yang et al. proposed automated multimodal indexing and retrieval system for German language lecture videos. Videos are indexed using both text and audio information. Authors have provided retrieving video text data in a structured format represented in terms of titles, subtitles and its content extracted from video slides using stroke width transforms and geometric information. Further, authors have proposed training open source CMU Sphinx model for German language and used this model for transcribing audio content into text. Most of the previous works utilized multiple modalities in the form of low level image features, visual text or audio content to segment and index lecture videos. However, the exploration of fusion of correlative and complementary features to improve retrieval performance for lecture videos is still remained as an open research issue.

Multimodal data fusion is the process of combining information from multiple sources and is becoming increasingly popular in the field of multimedia information indexing and

retrieval. One of the most popular approach for multimodal fusion is based on topic modeling and the most popular model being LDA. LDA is originated from natural language processing community. However it has gained great success in several fields including computer vision, multimedia retrieval and pattern recognition [19] [26]. In LDA models, documents are modeled as a multinomial distribution over topics, while each topic is modeled as multinomial distribution over words. We propose to use LDA algorithm to combine outputs of OCR and ASR and learn distribution of topics within video text and audio transcriptions. The LDA model uses document level word co-occurrence information and groups semantically related words into a single topic. The main objective of LDA models is to maximize the probability of observed data based on frequency of word distributions. However, these models have the tendency of sacrificing performance on rare topics in order to improve modeling distribution of frequently occurring words. Such models may create skewed impression when the input documents contain uneven word distributions especially in case of lecture videos where document collections are derived from complementary modalities such as video text and speech. We address this challenge by turning unsupervised topic modeling into semi-supervised by providing additional supervision information in the form of topics detected from instructional slides and we call them as seed words. Since OCR engines provide higher recognition rates, the seed words are obtained by using image text recognition and geometrical analysis. The model is thus encouraged to build word co-occurrence information for lecture video transcriptions generated from image text and audio modalities around the seed words in addition to regular words.

In the next section, the framework of proposed multimodal video indexing system is presented, the details of recognizing text from video and audio segments is described in the third section, in the fourth section the novel technique of integrating text and audio transcriptions using semi-supervised LDA is presented and in the final section, the experimental results are discussed.

II. MULTIMODAL VIDEO INDEXING FRAMEWORK

Fig. 1 shows the proposed multimodal architectural framework that takes input as lecture videos, extracts transcriptions using OCR and ASR engines and indexes using topic hierarchies obtained from LDA models. The proposed multimodal video indexing scheme comprises two separate channels for transcribing video text and audio contents into textual descriptions. The video text recognition pipeline segments videos into key frames based on slide transitions. The video key frames are preprocessed and text regions are localized before applying OCR for recognizing text information. In speech to text conversion pipeline, the slide transitions detected from video frames are utilized for detecting boundaries of audio segments. The CMU Sphinx ASR is applied on these audio segments to generate transcriptions. Recognition of text from preprocessed video key frames using current OCR technology provides average recognition rate of 90%. But, the

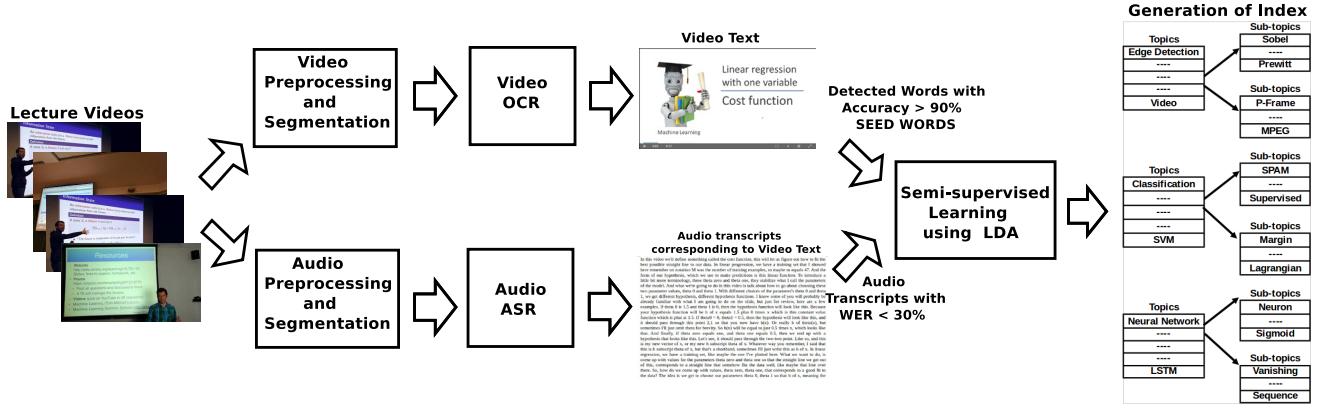


Fig. 1. LDA based multimodal fusion framework for segmentation and indexing of lecture videos.

text information on lecture slides will have limited content and may not satisfy requirements of the searches demanding comprehensive information about the topics and subtopics. In contrast to video text extraction, transcriptions obtained by applying ASR on real time audios are rich in terms of content and covers in-depth information about topics and subtopics. Nevertheless, the current ASR technologies have not achieved a low WER. In the proposed work, we utilize and integrate the complementary strengths of these two different modalities i.e high recognition rate of video text and content richness of audio transcriptions to improve retrieval performance of the indexing system.

The transcriptions from OCR and ASR engines provide a large collection of text documents which contain course content explained by the instructors through presentation slides. The LDA model divides these text document collections into semantically rich natural groups by exploiting co-occurrence relations between the words. It models each of these transcription documents as mixture of topics and topics as mixture of words. Such models are useful for extracting topic-subtopic information and index video recordings with semantic rich information. However, these models tend to scarify their performance on rare (but salient) topics especially in case of lecture videos where document collections are derived from video text (sparse but accurate) and speech (voluminous but high WER) information. We address this challenge by providing additional information in the form of topics and subtopics detected from instructional slides as seed words or supervision information. The model is thus trained and encouraged to build topic models around these seed words in addition to regular words. Further, as lecture video contents have inherent hierarchical structure, we address the problem of searching video contents at topic, subtopic or word levels by utilizing geometrical properties of video text and topic hierarchies detected by using LDA algorithm. We use the topic hierarchies derived from LDA to index lecture videos and attempted to address the problem of semantic gap by designing fusion based multimodal video indexing system which provides retrieval of video contents at topic and sub-

topic levels.

III. VIDEO TEXT AND SPEECH RECOGNITION

A. Video Segmentation and Text Recognition

Video segmentation involves representing a long lecture video with a set of unique and representative frames. We first convert videos into frames and sample frames at regular intervals $N \times W + 1$, with window size W and frame position N . Typically frame transition rate has smaller value in lecture videos compared to other video genres. We select length of W to be mean value obtained based on experimental evaluation. During text extraction from lecture videos, we separate frames comprising text information from non-slide frames. Several approaches [22] [8] were proven to be more effective for classification of video genre types using handcrafted features and machine learning techniques. However, handcrafted features are often non-scalable and fail to capture all the variations in the lecture slides. We propose to use Convolutional Neural Network (CNN) as binary image classifier for classification of video frames into slide and non-slide video frames. CNN uses alternating sequence of convolution and max-pooling layers and can capture hierarchical abstractions. Further, CNN is proven to be more suitable for recognition of objects and useful for differentiating lecture slides from non-lecture slides. The proposed CNN network for video segmentation operations comprises 2 layers of alternating sequence of convolution and max pooling layers followed by a dense fully connected layer. The frames classified by the CNN are used as inputs for the next phase for identification of unique frames and indexing of videos based on text information. Before recognizing text using OCRs, video frames are converted into grey scale followed by image binarization operation using Otsus algorithm. Edge gradients are obtained for each of these video frames using Sobel edge detection operator. The boundaries of text line segments are detected by applying Hough line transform on edge gradient maps. We use an open source Tesseract OCR engine to recognize text from each of these line segments. The slide transitions are detected by comparing text content of successive frames. The redundant frames are filtered by

comparing text contents based on position and height of text segments. The text information obtained is further filtered by applying post processing operations such as stop-word removal, stemming and lemmatization operations.

B. Generation of Audio Transcriptions

Transcribing audio content into text descriptions complement video OCR in retrieving relevant lecture videos. Building speech recognition system that converts audio into text is a challenging task due to variation in ascent, background noise, age, language and low quality recording systems. The speech recognition engines developed by using hand crafted features (MFCC, LPC etc) and classifier models (GMM and HMM) work well when the size of vocabulary is small and the models are less in number. Misclassification rate increases with the increase in size of vocabulary. Further, it becomes very challenging to build a scalable speech recognition engine that can adapt to a new user, ascent and variation in the recording systems. There are several open source ASRs (CMU SPHINX, Kaldi) and commercial (Microsoft, Google and Dragon) exist in the web repositories. We propose to use CMU Sphinx ASR engine for speech recognition as it is highly modular, portable and is implemented completely in Java programming language. Further, it is highly flexible and supports adaptation of the model for speech recognition with new language or change in accent. Generally lecture audio content is synchronized with the presentation slides, we propose to segment audio in accordance with the slide transitions detected during video segmentation. The CMU Sphinx ASR engine comprises 3 modules namely front end, decoder and knowledge base. Knowledge base in turn comprises acoustic, language and dictionary models. Input speech signals are first preprocessed and features are extracted in the front end module. Given a feature vector X , the ASR decodes it into text W by using Bayes formula.

Since our class room audio recordings belong to Indian English, we found a huge variation with respect pronunciation and ascent. We preferred to use US English models as default base models and try to adapt it to our class room recordings. Further, we used supervised model where the acoustic model is trained with known speaker recordings and their transcriptions. Once the model has been trained with new acoustic features, the existing acoustic model directory is updated with the new features.

IV. MULTI-MODAL TOPIC MODELING AND INDEXING USING LDA

LDA is a most popular unsupervised generative probabilistic topic modeling algorithm proposed by Blei et al. in 2003 [5] [14]. The basic principle of LDA is to represent documents as random mixture over latent topics and the topics in turn as random mixture over latent words. Let us consider a corpus C consisting of N documents with each document having M words, then LDA model uses following generative process to model corpus C .

- Let t denotes T number of topics $\{1, 2, 3, \dots, T\}$. Select multinomial distribution ϕ_t for each topic t from a Dirichlet distribution $\text{Dir}(\beta)$
- Similarly select multinomial distribution θ_d for each document $\{1, 2, 3, \dots, N\}$ from a Dirichlet distributions $\text{Dir}(\alpha)$
- For each word w_n in document d where $n \in 1, 2, \dots, d_N$
 - select a topic t_n from multinomial distribution of θ_d
 - select a word w_n from multinomial distribution of ϕ_{t_n}

The words from slide text and audio files are observed variables. The hyper parameters (ϕ, θ) and latent variables (α, β) are inferred by maximizing the probability distribution of observed data D as follows

$$P(D|\alpha, \beta) = \prod_{i=1}^N \int P(\theta_d|\alpha)(f)d\theta_dd_p \quad (1)$$

where f is given by

$$f = \sum_{n=1}^M P(z_{d_n}|\theta_d)P(w_{dn}|z_{dn}, \phi)P(\phi|\beta) \quad (2)$$

When the LDA model is fit to the text outputs of OCR and ASR engines, model utilizes the co-occurrence relations between the words and tries to group semantically related words. Due to higher WER of ASR engines LDA models have the tendency to group either semantically correlated words under a wrong topic or grouping incorrect words into a right topic. On the other hand fitting model only to OCR engine outputs might result into accurate modeling of topics to word distributions, but results will not provide in depth information about topics and subtopics. As we have confidence over text recognition outputs, we utilize these words as seed words and try to guide LDA model to group around these seed topics in addition to regular topics.

The proposed semi-supervised LDA algorithm comprises two models where topics are represented as multinomial distribution over words in the first model and documents are defined as multinomial distribution over topics in the second model. In the conventional LDA, each topic is represented by a multinomial distribution ψ . In semi-supervised LDA model, we define each topic as the convex combination of multinomial distributions of regular topics ψ_r and seed topics ψ_s . The parameter π is the probability value to pick words from either seeded topics or regular topics. The seed topics are set according to the words detected by Tesseract text recognition engine. Therefore from the first model, multinomial distributions are generated for both regular and seeded topics. Then, topics are generated belonging to either seeded or regular topic distributions and words are generated for the selected topic distribution. The algorithm 1 describes the generative story of proposed semi-supervised LDA.

V. EXPERIMENTAL RESULTS

The experimental video dataset comprises 500 lecture videos collected from Coursera, NPTEL and KLETU [1]

Algorithm 1 semi-supervised LDA Algorithm

```

1: for  $k = 1 \rightarrow T(\text{topics})$  do
2:   Draw  $\phi_k^r$  for regular topics from  $\text{Dir}(\beta_r)$ 
3:   Draw  $\phi_k^s$  for seed topics from  $\text{Dir}(\beta_s)$ 
4:   select  $\pi_k$  from distributions  $Beta(1, 1)$ 
5: end for
6: for  $s = 1 \rightarrow S(\text{seedSets})$  do
7:   select group-topic distribution  $\psi_s$  from  $\text{Dir}(\alpha)$ 
8: end for
9: for  $d = 1 \rightarrow N(\text{Documents})$  do
10:  select a binary vector  $\bar{b}$  of length  $S$ 
11:  select a document-group distribution  $\eta^d$  from  $\text{Dir}(\tau\bar{b})$ 
12:  select a group variable  $g$  from multinomial distribution
     $\eta^d$ 
13:  select  $\theta_d$  from multinomial distribution  $\psi^g$ 
14:  for  $i = 1 \rightarrow n(\text{tokens})$  do
15:    Choose a topic  $z_i$  from multinomial distributions
        of  $\theta_d$ .
16:    Choose an indicator  $x_i$  from  $Bern(\pi_{z_i})$ 
17:    if  $x_i == 0$  then
18:      Select a word  $w_i$  from  $Mult(\phi_{z_i}^r)$ 
19:    else
20:      Select a word  $w_i$  from  $Mult(\phi_{z_i}^s)$ 
21:    end if
22:  end for
23: end for

```

TABLE I
PERCENTAGE IMPROVEMENT IN MAP USING MULTIMODAL FUSION

Modality	Coursera	NPTEL	Classroom
Uncorrected			
Text vs Speech+Text	15.93	20.78	28.48
Speech vs Speech+Text	11.17	11.87	2.17
Corrected			
Text vs Speech+Text	10.93	6.05	6.82
Speech vs Speech+Text	3.17	8.47	2.44

classrooms. The proposed model is evaluated on videos comprising instructors explaining topics using presentation slides. For validation purpose, we also collected audio transcripts which are available with the downloaded videos. For the videos which are not associated with speech transcripts, we generated it manually. The TensorFlow and Keras libraries are used to build CNN classifier model. We use Gensim Python API for implementing semi-supervised LDA fusion based topic models. Since the performance of LDA models largely depends on the volume of training data, we used text data extracted from online books from the Engineering stream for training LDA models. The model is then fit to the text transcriptions obtained from lecture videos. We build separate index tables by utilizing topic hierarchies generated from unimodal and multimodal inputs. The performance of different index tables are evaluated for 200 queries by using Mean Average Precision (MAP) measure. The Table I depicts the percentage improvement in MAP of multimodal fusion based indexing system over unimodal indexing.

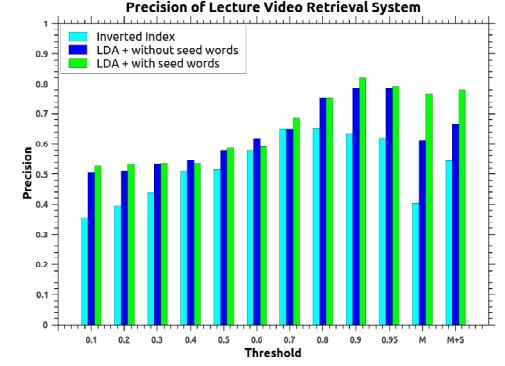


Fig. 2. Improvement in precision results obtained with the addition of topic hierarchies to the index table. The topic hierarchies are derived from LDA and semi-supervised LDA models.

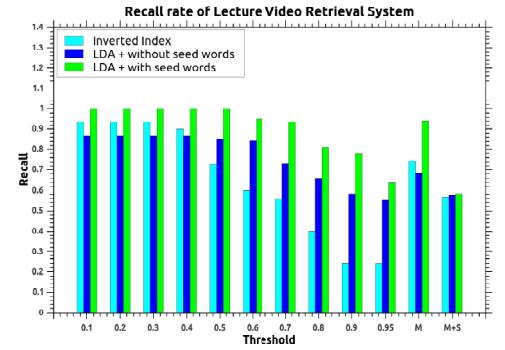


Fig. 3. Recall rate results obtained with the addition of topic hierarchies to the index table.

Further, in order to evaluate the flexibility and percentage improvement of hierarchical indexing and retrieval system, we follow the method used by Myaeng S et al. [15]. Relevance judgments for the retrieval results are made by course instructors who are specialized in the course. For each query, the experts first select a video shot which is considered as most relevant. Then the relevant topic-subtopic information in that document against this query are judged and selected by the experts without the knowledge of the targeting systems. We evaluate the performance of proposed flexible hierarchical indexing and retrieval system by setting various thresholds. Thresholds signify fractions of the total retrieved results returned to users. We use both fixed thresholds from 0.1 to 0.9 plus 0.95 and two dynamic thresholds. The fixed thresholds range from 0.1 to 0.9 and two dynamic thresholds are set by using *mean* and *stddev*. The dynamic thresholds *mean* and *stddev* are computed by using the rank values of the retrieved records. The index tables constructed for topics and subtopic information are compared with TF-IDF based inverted indexing system. The charts shown in Figures 2 and 3 illustrates the improvement in precision and recall rates obtained using LDA models compared to inverted index based techniques. Following Fig. 4 we observe an average improvement of 29.09% and 44.49% in retrieval performance

without using seed words and with the addition of seed words respectively.

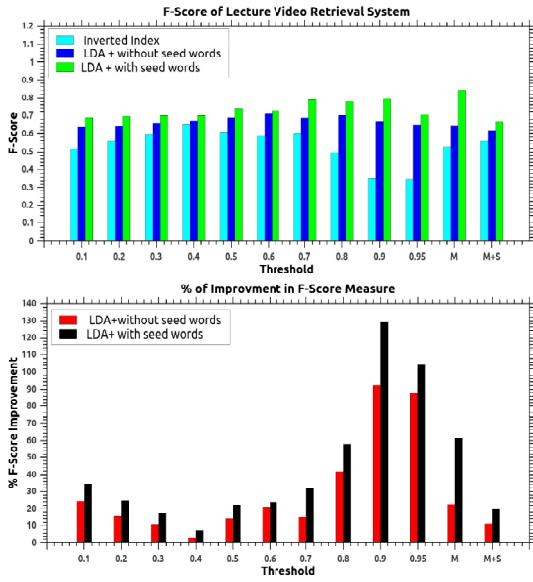


Fig. 4. Improvement in the F-Score rate with the addition of topic hierarchies derived from LDA and semi-supervised LDA models.

VI. CONCLUSION

Indexing lecture videos directly with the transcriptions obtained from video presentation slides and audio content results into poor retrieval performance due to data sparsity and high WER respectively. We addressed this problem by integrating the complementary strengths of video text and speech modalities by proposing a novel idea of fusion using semi-supervised LDA algorithm. The LDA model exploits word co-occurrence relationships existing between speech and slide text information for deriving latent semantic relations and provides indexing information in the form of topic hierarchies. The proposed model is evaluated on Coursera, NPTEL and KLETU lecture videos. The experimental results prove the efficacy of employing fusion based models for indexing lecture videos. We achieved an improvement in F-Score of 44.49% for fusion based indexing models compared with indexing using unimodal approaches. In future, proposed work can be extended to consider handwritten text and support multiple video genres.

REFERENCES

- [1] Kle technological university, hubballi. <https://www.kletech.ac.in/>. Accessed: 2018-09-15.
- [2] Speech based audio indexing. <http://speechfind.utdallas.edu/>. Accessed: 2018-03-30.
- [3] G. D. Abowd. Classroom 2000: An experiment with the instrumentation of a living educational environment. *IBM Systems Journal*, 38(4):508–530, 1999.
- [4] Christopher Alberti, Michiel Bacchiani, Ari Bezman, Ciprian Chelba, Anastassia Drofa, Hank Liao, Pedro Moreno, Ted Power, Arnaud Sahuguet, Maria Shugrina, and Olivier Siohan. An audio indexing system for election video material. In *Proceedings of ICASSP*, pages 4873–4876, 2009.
- [5] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.
- [6] C. Y. Chiu, P. C. Lin, S. Y. Li, T. H. Tsai, and Y. L. Tsai. Tagging webcast text in baseball videos by video segmentation and text alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(7):999–1013, July 2012.
- [7] Matthew Cooper. Presentation video retrieval using automatically recovered slide and spoken text. 8667, 03 2013.
- [8] Ling-Yu Duan, Min Xu, Qi Tian, Chang-Sheng Xu, and J. S. Jin. A unified framework for semantic shot classification in sports video. *IEEE Transactions on Multimedia*, 7(6):1066–1083, Dec 2005.
- [9] Kartik Dutta, Minesh Mathew, Praveen Krishnan, and C. V. Jawahar. Localizing and recognizing text in lecture videos. In *ICFHR*, 2018.
- [10] G. Heigold, H. Ney, R. Schlueter, and S. Wiesler. Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance. *IEEE Signal Processing Magazine*, 29(6):58–69, Nov 2012.
- [11] Mahima Joshi. Evaluation of speech and text-based indexing for classroom lecture videos. *Thesis, Ph.D.*, 2015.
- [12] W. Kim and C. Kim. A new approach for overlay text detection and extraction from complex video scene. *IEEE Transactions on Image Processing*, 18(2):401–411, Feb 2009.
- [13] Rainer Lienhart. *Video OCR: A Survey And Practitioner's Guide*, pages 155–183. Springer US, Boston, MA, 2003.
- [14] Jon D. McAuliffe and David M. Blei. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. Curran Associates, Inc., 2008.
- [15] Sung Hyon Myaeng, Don-Hyun Jang, Mun-Seok Kim, and Zong-Cheol Zhoo. A flexible model for retrieval of sgml documents. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '98*, pages 138–145, 1998.
- [16] N. V. Nguyen, M. Coustaty, and J. M. Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *2014 22nd International Conference on Pattern Recognition*, pages 2667–2672, Aug 2014.
- [17] Nhu Van Nguyen, Mickal Coustaty, and Jean-Marc Ogier. Multi-modal and cross-modal for lecture videos retrieval. In *Proceedings of the 2014 22Nd International Conference on Pattern Recognition, ICPR '14*, pages 2667–2672, Washington, DC, USA, 2014. IEEE Computer Society.
- [18] Jun Ogata and Masataka Goto. Podcastle: A spoken document retrieval system for podcasts and its performance improvement by anonymous user contributions. In *Proceedings of the Third Workshop on Searching Spontaneous Conversational Speech, SSCS '09*, pages 37–38, New York, NY, USA, 2009. ACM.
- [19] Stephen Roller and Sabine Schulte im Walde. A multimodal lda model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1146–1157, Seattle, WA, October 2013.
- [20] Tayfun Tuna. Automated lecture video indexing with text analysis and machine learning. *Thesis, University of Houston*, 2015.
- [21] Tayfun Tuna, Jaspal Subhlok, Lecia Barker, Shishir Shah, Olin Johnson, and Christopher Hovey. Indexed captioned searchable videos: A learning companion for stem coursework. *Journal of Science Education and Technology*, pages 82–99, 2017.
- [22] Yao Wang, Zhu Liu, and Jin-Cheng Huang. Multimedia content analysis—using both audio and visual clues. *IEEE Signal Processing Magazine*, 17(6):12–36, Nov 2000.
- [23] H. Yang and C. Meinel. Content based lecture video retrieval using speech and video text information. *IEEE Transactions on Learning Technologies*, 7(2):142–154, April 2014.
- [24] Sheng yi Kong, Miao ru Wu, Che-Kuang Lin, Yi-Sheng Fu, and Lin-Shan Lee. Learning on demand - course lecture distillation by information extraction and semantic structuring for spoken documents. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4709–4712, 2009.
- [25] D. Yu, L. Deng, and F. Seide. The deep tensor neural network with applications to large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):388–396, Feb 2013.
- [26] Y. Zheng, Y. Zhang, and H. Larochelle. Topic modeling of multimodal data: An autoregressive approach. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1370–1377, June 2014.

Emotion Recognition from Varying Length Patterns of Speech using CNN-based Segment-Level Pyramid Match Kernel based SVMs

Shikha Gupta^{*1}, Kishalaya De^{*2}, Dileep Aroor Dinesh ¹ and , Veena Thenkanidiyoor³

¹School of Computing and EE, Indian Institute of Technology, Mandi, H.P., India

²Department of ECE, Manipal Institute of Technology, Udupi, Karnataka, India

³Department of CSE, National Institute of Technology, Goa, Ponda, Goa, India

shikha_g@students.iitmandi.ac.in, kishalaya97@gmail.com, addileep@iitmandi.ac.in,
veenat@nitgoa.ac.in

Abstract—Convolutional Neural Networks (CNNs) and its variants have achieved impressive performance when used for different speech processing tasks like spoken language identification, speaker verification, speech emotion recognition, etc. Conventionally, CNNs for speech applications consider input features from fixed duration speech segments as input. In this work, we attempt to consider features from complete speech signal as input to CNN. We propose to use spatial pyramid pooling (SPP) layer in CNN architecture to remove the fixed length constraint and to consider features from varying length speech signals as input to CNN for an end to end training. Proposed architecture also results in varying size set of feature maps from convolution layer. Further, we propose novel CNN-based segment-level pyramid match kernel (CNN-SLPMK) as dynamic kernel between a pair of varying size set of feature maps for the classification framework using support vector machines (SVMs) based classifier. We demonstrate that our proposed approach achieves comparable results with state-of-the-art techniques for speech emotion recognition task.

I. INTRODUCTION

Speech emotion recognition (SER) is one of the challenging tasks in speech processing [1]. Knowing the emotion embedded in the speech is important for natural and seamless interaction of humans with many applications of automated speech query dialog systems. Same is also evident from the emerging idea of automated real time call centers. However, recognizing emotion from speech signal is still a very difficult problem as effective feature representation and building the efficient models are open questions [1], [2].

Traditional SER methods involve first extracting energy based low level features such as Mel frequency cepstral coefficients (MFCC) [3] from speech utterances and then generating global dictionary by applying clustering techniques over training set. This dictionary is further used for converting variable length input example to fixed dimensional feature vector for the classification using Gaussian mixture model (GMM) or support vector machine (SVM) based classifiers [4], [5], [6].

^{*}Both authors contributed equally to this work.

In the recent years, CNNs have become popular for their applicability to a wide range of tasks in image domain [7], [8], [9] and speech processing [11], [12]. The features learned using CNNs have achieved remarkable performance due to their capability of modeling temporal local correlations and reducing translational variations [13], [12]. The main issue with traditional CNNs is that they require fixed size input for training and testing. The input examples size in real world are varying in nature, for e.g, in vision domain, scene images significantly vary from one another in size and aspect ratio [15], [10]. Similarly in speech domain speech utterances are variable length in nature [16], [17]. In this work, we propose considering original varying size speech signal features as an input to the CNN for end to end training instead of fixed length context segments. This reduces the work of converting variable length speech signal to fixed length context segments and at the same time enhances model's performance.

In this work, we propose two approaches to remove fixed size constraint in CNN. In the first approach we propose to modify the architecture of CNN in such a way that, feature matrix from complete speech signal is considered as input. This results in varying size set of feature maps for each speech signal. Inspired from [15] on visual recognition, we propose to use spatial pyramid pooling (SPP) layer to map varying size set of feature maps onto a fixed length representation for end-to-end training of the CNN. In [15], SPP-layer is added to AlexNet [13] and modified architecture is named as SPP-Net. SPP-Net computes the feature maps from the entire image (without resizing) and then pools feature map in spatial regions to generate fixed-length representations for training the network end-to-end.

In the second approach, we focus on using the CNN architecture which allow processing the feature representation of varying length speech signal which is essential during training and testing. Once the network is trained, it can be used to recognize emotion from varying length speech signal. It is found in different image classification tasks that SVM-based classifiers using features extracted from last convolutional layer are very effective [18], [19], [10]. The convolutional

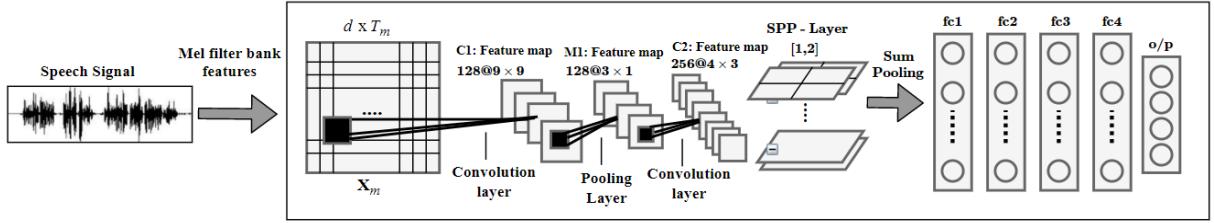


Fig. 1. Proposed CNN architecture with SPP-layer. Here, T_m is the number of frames in speech signal and d is the dimension of feature vector representing a frame.

layers are the indispensable part of CNN and responsible for generating discriminative features. So, we focus on considering the set of feature maps from last convolutional layer and design a suitable kernel to improve the performance of SER using SVM-based classifier. Since the CNN in this work, consider varying length speech signal features as input, varying size set of feature maps are obtained at the last convolutional layer. The varying size feature maps are obtained either by considering a pre-trained CNN for SER or by training CNN with SPP layer. We propose to build a novel CNN-based segment-level pyramid match kernel (CNN-SLPMK) between a pair of varying size set of feature maps. The varying size feature maps for two speech signals are further temporally divided into segments and sum-pooled at each level of pyramid. These temporally pooled feature maps are l_1 -normalized. Then, we propose to compute a matching score between normalized vector representation of pair of speech signals at that level of pyramid. The matching score at different levels is combined to get CNN-SLPMK between a pair of varying length speech signals.

The key contributions of this paper are:

- Introducing SPP-layer [15] in between last convolutional layer and first dense layer so that variable size convolutional feature map of speech signal can be converted into fixed length representation.
- A simple CNN architecture with two convolution layers followed by SPP-layer and dense layers for end-to-end training the network for SER.
- CNN-based segment-level pyramid match kernel (CNN-SLPMK) to find the similarity score between a pair of varying length speech signals.
- Effectiveness of the varying length feature maps obtained either by considering a pre-trained CNN or by CNN with SPP layer is verified on FAU-AEC [17] and EMO-DB [16] SER datasets using CNN-SLPMK.

The remaining paper is organized as follows: Section II discuss about conventional CNN for speech application. Section III presents the modified CNN architecture with SPP-layer. The proposed CNN-SLPMK for varying size set of feature maps is presented in Section IV. The details about datasets, training of network, feature extraction, classification framework and experimental results on speech emotion recognition are given in Section V. Finally, we conclude the paper

and discuss future work in Section VI.

II. CNN FOR SPEECH APPLICATIONS

Conventional CNNs for different speech tasks [20], [21], [12] take fixed size feature matrix of size $d \times l_c$ corresponding to a context segment as input. Here, d is the dimension of feature vectors corresponding to a frame and l_c is the number of frames considered in a context segment. Consider a speech signal of an utterance with T frames which is represented as a set of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t, \dots, \mathbf{x}_T\}$, where \mathbf{x}_t is a d -dimensional feature vector for frame t . For every \mathbf{x}_t in set of feature vector \mathbf{X} , l feature vectors to the left and r feature vectors to the right is considered for generating a context segment of fixed size $d \times l_c$ where, $l_c = l + r + 1$.

In this work, we have considered a baseline CNN architecture explained in [21], which include two convolutional layers and a max-pool layer. The first convolutional layer has 128 filters each of size 9×9 followed by 3×1 max-pool layer and second convolutional layer has 256 filters each of size 4×3 . Following the second convolutional layer, the baseline CNN architecture includes 4 fully connected (fc) layers with 2048 neurons in each layer followed by a softmax output layer. We focus on using the CNN architecture that allow for processing the feature representation of varying length speech signals instead of fixed size context segments. Varying length input to the CNN architecture results in varying size set of feature maps at the last convolutional layer. To handle this, we propose two approaches for the classification of SER task. First, we propose to use spatial pyramid pooling (SPP) layer [15] in between convolutional layer and fc layer to map varying size set of feature maps onto a fixed length representation for end-to-end training. Second, we propose novel CNN-based segment level pyramid match kernel as dynamic kernel using SVMs to compute the similarity score between pair of varying size set of feature maps.

III. CNN ARCHITECTURE WITH SPP-LAYER

In this work, we propose to add a SPP-layer in between last convolutional layer and the first fc-layer. The SPP-layer sum-pools the varying length convolutional layer feature maps at two different levels to convert them into fixed length vector. In the first level complete convolutional feature map is considered and sum-pool is applied to obtain a fixed length vector. In the

second level feature map is temporally divided into 4 segments and sum pooling is applied in respective part for converting the variable size deep feature maps to fixed size. The fixed length vectors obtained in each level are concatenated to form a fixed length supervector. This fixed length supervector is further passed onto fully connected layer for end-to-end training. The proposed CNN architecture with SPP-layer is shown in Figure 1. The speech signal of an utterance with T_m frames is represented as a set of feature vectors $\mathbf{X}_m = \{\mathbf{x}_{m1}, \mathbf{x}_{m2}, \dots, \mathbf{x}_{mT_m}\}$, where \mathbf{x}_t is a d -dimensional feature vector for frame t . When this $d \times T_m$ feature matrix is given as input to CNN, set of feature maps¹ $\mathcal{X}_m \in \mathbb{R}^{m_p \times m_q \times f}$ is generated from last convolutional pooling layer of CNN.

It is found in many image understanding tasks that SVM-based classifiers using features learned from last convolutional layer are very effective [19]. To build a discriminative classifier like SVM for varying size feature maps obtained from last convolutional layer, a suitable kernel is required. In the next Section, we propose CNN-SLPMK for the varying size set of feature maps.

IV. CNN-BASED SEGMENT-LEVEL PYRAMID MATCH KERNEL

Inspired by the work of [22], [23], we propose novel CNN-based segment-level pyramid match dynamic kernel to handle variable size feature maps for speech emotion recognition task. CNN-SLPMK measures the similarity score between two speech signal of same or different lengths. The process of computing CNN-SLPMK is illustrated in Figure 2. Let $\mathbf{X}_m = \{\mathbf{x}_{m1}, \mathbf{x}_{m2}, \dots, \mathbf{x}_{mT_m}\}$ and $\mathbf{X}_n = \{\mathbf{x}_{n1}, \mathbf{x}_{n2}, \dots, \mathbf{x}_{nT_n}\}$ be the two set of d -dimensional feature vectors corresponding to two speech signals of length T_m and T_n respectively. When these $d \times T_m$ and $d \times T_n$ variable size feature matrices are given as input to pre-trained CNN using conventional method [21] or CNN with SPP-layer, varying size set of feature maps $\mathcal{X}_m = \{\hat{\mathbf{x}}_{m1}, \dots, \hat{\mathbf{x}}_{mi}, \dots, \hat{\mathbf{x}}_{mf}\} \in \mathbb{R}^{m_p \times m_q \times f}$ and $\mathcal{X}_n = \{\hat{\mathbf{x}}_{n1}, \dots, \hat{\mathbf{x}}_{ni}, \dots, \hat{\mathbf{x}}_{nf}\} \in \mathbb{R}^{n_p \times n_q \times f}$ are generated from last convolutional pooling layer of f filters. Here, $\hat{\mathbf{x}}_{mi}$ and $\hat{\mathbf{x}}_{ni}$ are the i^{th} feature map for m^{th} and n^{th} speech utterance respectively. Size of feature maps in a set corresponding to one speech signal is different from that of another speech signal, because size of feature map is dependent on the input feature matrix size. Computation of CNN-SLPMK is described in details in Algorithm 1.

CNN-SLPMK operates over pair of set of features maps with L pyramid levels. At level l , each of the feature map $\hat{\mathbf{x}}_{mi}$ and $\hat{\mathbf{x}}_{ni}$, $\forall i \in f$, are temporally divided into 2^{2l} blocks. Sum pooling is applied over each block of every feature map which results in feature vector \mathbf{X}_m^l and $\mathbf{X}_n^l \in \mathbb{R}^{(2^{2l} \times f) \times 1}$. These feature vectors are further normalized using ℓ_1 - normalization technique. Similarity score between a pair of feature maps set is obtained by computing intermediate matching score S_l using histogram intersection function of Equation 2. Final matching

¹ $\mathcal{X}_m = \{\mathbf{x}_{m1}, \dots, \mathbf{x}_{mi}, \dots, \mathbf{x}_{mf}\}$; where $\mathbf{x}_{mi} \in \mathbb{R}^{m_p \times m_q}$ is the i^{th} feature map, f = number of filters in last convolutional layer of CNN.

Algorithm 1 CNN-based segment-level pyramid match kernel $K_{\text{CNN-SLPMK}}(\mathcal{X}_m, \mathcal{X}_n)$

Inputs:

- (i) Varying size set of feature map,
 $\mathcal{X}_m = \{\hat{\mathbf{x}}_{m1}, \dots, \hat{\mathbf{x}}_{mi}, \dots, \hat{\mathbf{x}}_{mf}\}$ where, $\hat{\mathbf{x}}_{mi} \in \mathbb{R}^{m_p \times m_q}$
 $\mathcal{X}_n = \{\hat{\mathbf{x}}_{n1}, \dots, \hat{\mathbf{x}}_{ni}, \dots, \hat{\mathbf{x}}_{nf}\}$ where, $\hat{\mathbf{x}}_{ni} \in \mathbb{R}^{n_p \times n_q}$
- (ii) L : total number of pyramid levels.

Procedure:

- 1: **for** $l = 0$ **to** $L - 1$ **do**
- 2: At level l , divide each feature map of \mathcal{X}_m and \mathcal{X}_n into 2^{2l} blocks.
- 3: Compute $\mathbf{X}_m^l \in \mathbb{R}^{f2^{2l} \times 1}$ and $\mathbf{X}_n^l \in \mathbb{R}^{f2^{2l} \times 1}$ after applying sum pooling over each block of feature maps.
- 4: ℓ_1 - normalize the generated feature vectors \mathbf{X}_m^l and \mathbf{X}_n^l .

$$\hat{\mathbf{X}}_m^l = \frac{\mathbf{X}_m^l}{\sum_{j=1}^{f2^{2l}} \mathbf{X}_m^l}, \quad \hat{\mathbf{X}}_n^l = \frac{\mathbf{X}_n^l}{\sum_{j=1}^{f2^{2l}} \mathbf{X}_n^l} \quad (1)$$

- 5: Compute level-wise matching score using histogram intersection function as:

$$S_l = \sum_{j=1}^{f2^{2l}} \min(\hat{\mathbf{X}}_m^l[j], \hat{\mathbf{X}}_n^l[j]) \quad (2)$$

- 6: **end for**

- 7: Compute final similarity score between \mathcal{X}_m and \mathcal{X}_n as:

$$K_{\text{CNN-SLPMK}}(\mathcal{X}_m, \mathcal{X}_n) = \sum_{l=0}^{L-2} \frac{1}{2^{(L-l-1)}} (S_l - S_{l+1}) + S_{L-1} \quad (3)$$

Outputs:

- (i) $K_{\text{CNN-SLPMK}}(\mathcal{X}_m, \mathcal{X}_n)$.
-

score $K_{\text{CNN-SLPMK}}$ is obtained by weighted combination of level-wise similarity score using Equation 3.

V. EXPERIMENTAL STUDIES

In this section, details of databases used, feature representation and effectiveness of the proposed approaches is studied for speech emotion recognition tasks using SVM-based classifiers.

A. Databases

In this work, the German FAU Aibo emotion corpus (FAU-AEC) [17] and the Berlin emotional database (Emo-DB) [16] are used for conducting the experiments on speech emotion recognition task.

- **FAU-AEC:** This dataset comprises of German speech emotions recording of 51 children between the age of 10 to 13 of two different schools interacting with robot named Aibo, within the setup explained in [17]. In total there were 48,401 words, which were clubbed into semantically meaningful chunks (collectively called speech utterance), with a statistic of 2.66 words per chunk on average. We have considered four super classes

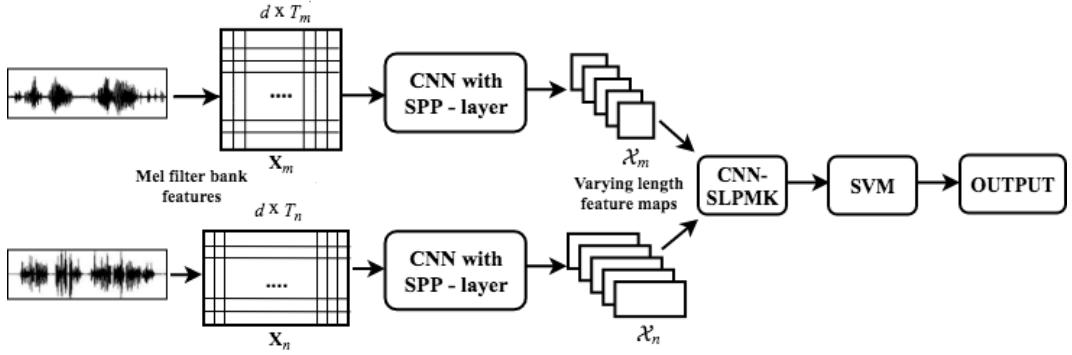


Fig. 2. Illustration of computation of CNN-SLPMK between two speech signals of different length. Here, varying size set of deep feature maps \mathcal{X}_m and \mathcal{X}_n are obtained from last convolutional layer of CNN architecture with SPP-layer as in Figure 1.

of emotions – namely, *anger*, *emphatic*, *neutral*, and *motherese*. This dataset includes a total of 4,543 speech utterances. Classification is performed at speech utterance level. Speaker independence was guaranteed by using the data of one school as training and the data of other school as testing. The speaker-independent speech emotion recognition accuracy presented in this study for the FAU-AEC is the average classification accuracy along with 95% confidence interval obtained for 5-fold stratified cross validation. Thus in each fold, the training set includes about 3,000 and the test set includes about 1,500 speech utterances.

- *Emo-DB*: This dataset comprises of seven emotional classes – namely, A: *anger* (127), B: *boredom* (79), D: *disgust* (38), F: *anxiety / fear* (127), H: *happiness* (64), S: *sadness* (53) and N: *neutral* (78) as speaker emotions. Spoken sentences were pre-decided which included 10 German emotionally neutral sentences uttered by five male and five female professional actors. We have considered 80% of the utterances for training and the remaining for testing. The multi-speaker speech emotion recognition accuracy presented in this work for the Emo-DB is the average classification accuracy along with 95% confidence interval obtained for 5-fold stratified cross-validation.

B. Experimental setup

We have considered a baseline CNN architecture proposed in [21] which is also explained in Section III. The frames of size 20 ms with shift of 10 ms from the speech signal of an utterance are used for feature extraction. Every frame is represented using a 40-dimensional Mel filter bank coefficients. The frame level features thus extracted from the complete utterance are combined in the matrix form and given as input to the CNN. Size of utterance varies from one another so as the number of frames and size of input matrix. Hence the output at convolutional layer is also varying feature maps from one example to the other.

We propose to add two-level SPP-layer between last convolutional layer and the first fc-layer when the speech signal

with original size is given as input. The first level performed a sum-pooling over the feature maps. The second level divide the feature maps into four blocks, followed by sum pooling over individual blocks. The SPP layer converts the variable size feature maps to fixed size representation. The fixed size output of the SPP layer is given as input to the dense layer network. The dense layer neural network consists of 4 layers, each with 2048 hidden units. In the literature, several state-of-the-art deep architectures for CNN are available [12], [20]. However, we have considered only the baseline architecture as the size of databases we have considered for the study are relatively small. In principle, SPP-layer can be added to any state-of-the-art deep CNN architecture for allowing variable size original speech signal features as input and converting variable size set of feature maps from convolutional layer to fixed size representation for end to end training.

We first study the effectiveness of the proposed CNN architecture with SPP-layer for speech emotion recognition and compare with that of the baseline CNN architecture. For baseline CNN, feature matrix of size 40×21 is considered as input. Here, 40 is the dimension of feature corresponding to each frame and 21 is the number of frames in the context segment. A context segment is considered for every feature vector (i.e frame) by considering 10 frames to the left and 10 frames to the right. For CNN architecture with SPP-layer, a feature matrix of size $40 \times T_m$ is considered as input. Here, T_m is the number of frames in the m^{th} speech utterance. The CNNs are trained by minimizing the cross entropy as objective function. The value of learning rate and dropout are chosen empirically. In this work the best results are observed for the value 10^{-4} as learning rate and 0.25 as value for dropout in almost all the cases. Next we also study the effectiveness of the CNN-SLPMK based SVM classifier for speech emotion recognition task. The feature maps at the output of the convolution layer were chosen as the best representation of the input speech signal and given to the kernel for further classification. We consider, LIBSVM [24] tool to build the SVM-based classifiers. In this study, one-against-the-rest approach is considered for 7-class and 4-

class speech emotion recognition tasks. The value of trade-off parameter, C in SVM is chosen empirically as 10^2 .

C. Experimental Results

Table I compares the accuracies for the speech emotion recognition task obtained using baseline CNN, proposed CNN with SPP-layer and SVM-based classifier using proposed CNN-SLPMK. It is seen that the proposed CNN with SPP-layer performs marginally better than that of the baseline CNN for speech emotion recognition using FAU-AEC dataset. It is also observed that SVM-based classifier using CNN-SLPMK performed significantly better than that of the baseline CNN and CNN with SPP-layer for speech emotion recognition using FAU-AEC. However, the proposed CNN with SPP-layer and SVM with CNN-SLPMK performed poorer than that of the baseline CNN for speech emotion recognition using Emo-DB. The main reason is that the Emo-DB is relatively a very small dataset as compared to that of FAU-AEC.

TABLE I
COMPARISON OF CLASSIFICATION ACCURACY (CA) (IN %) WITH 95% CONFIDENCE INTERVAL FOR THE BASELINE CNN, PROPOSED CNN WITH SPP-LAYER AND SVM-BASED CLASSIFIER USING THE PROPOSED CNN-SLPMK FOR SPEECH EMOTION RECOGNITION TASK. THE HIGHEST ACCURACY OF EACH COLUMN IS MARKED IN BOLD.

Classification model	Emo-DB	FAU-AEC
Baseline CNN	86.12 ± 0.16	61.75 ± 0.12
CNN with SPP-layer	80.76 ± 0.13	63.85 ± 0.11
SVM using CNN-SLPMK	82.33 ± 0.15	66.43 ± 0.11
SVM using CNN-SLPMK obtained from pre-trained CNN	91.23 ± 0.13	-

Though we have considered only four superclass of emotions, FAU-AEC include all the emotions that are present in Emo-DB. In the next experiment we considered CNN with SPP-layer built for FAU-AEC dataset as pre-trained model. The weights of this pre-trained CNN model are kept fixed without fine-tuning. We have passed the feature matrix of varying length speech signals from Emo-DB as input to this CNN model and extracted varying size set of feature maps from last convolutional layer. The CNN-SLPMK for Emo-DB is now computed using the varying size feature maps obtained from pre-trained CNN model. The performance of this CNN-SLPMK based SVM classifier for speech emotion recognition using Emo-DB is given in Table I. It is seen from Table I that, the SVM-based classifier with CNN-SLPMK obtained from pre-trained CNN is performed significantly better than baseline CNN. This shows the effectiveness the CNN-SLPMK when the pre-trained model is available. Another advantage is that, computation complexity in computing CNN-SLPMK reduces significantly as we do not need to train the CNN model from the scratch.

Table II compares the accuracies for speech emotion recognition task obtained using the Bayes classifier using GMM and SVM-based classifiers using the state-of-the-art dynamic kernels, baseline CNN, proposed CNN with SPP-layer and

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY (CA) (IN %) WITH 95% CONFIDENCE INTERVAL OF THE SVM-BASED CLASSIFIER USING PROPOSED CNN-SLPMK, PROPOSED CNN WITH SPP-LAYER WITH THE BAYES CLASSIFIER USING GMM, SVM-BASED CLASSIFIERS USING STATE-OF-THE-ART DYNAMIC KERNELS AND BASELINE CNN FOR SPEECH EMOTION RECOGNITION TASK. THE HIGHEST ACCURACY OF EACH COLUMN IS MARKED IN BOLD.

Classification model	Emo-DB	FAU-AEC
MLGMM [25]	66.81 ± 0.44	60.00 ± 0.13
Adapted GMM [25]	79.48 ± 0.31	61.09 ± 0.12
Baseline CNN	86.12 ± 0.16	61.75 ± 0.12
CNN with SPP-layer (Ours)	80.76 ± 0.13	63.85 ± 0.11
SVM using GMMIMK [25]	85.62 ± 0.29	62.48 ± 0.07
SVM using FK [25]	87.05 ± 0.24	61.54 ± 0.11
SVM using PSK [25]	87.46 ± 0.23	62.54 ± 0.13
SVM using CNN-SLPMK (Ours)	91.23 ± 0.13	66.43 ± 0.11

SVM classifier using the proposed CNN-SLPMK. In this study, Bayes classifier using GMM is build using two ways, first parameters are estimated using the maximum likelihood method (MLGMM) and second the parameters of the UBM are adapted to the class specific data (adapted GMM) [4] with effective number of Gaussian components as 256. Diagonal covariance matrices are used to built the GMM in both the cases. SVM-based classifier with dynamic kernels namely, GMM based intermediate matching kernel (GMMIMK), Fisher kernel (FK) and probabilistic sequence kernel (PSK) is used for comparison. The details of the experiments and the best values for the parameters are considered as in [25] and [23]. GMMIMK is based on intermediate matching kernel (IMK) proposed in [26]. IMK is computed by matching two sets of varying length feature vectors using a set of virtual feature vectors. In literature virtual feature vectors are considered as codebook based cluster centers or GMM based cluster center. In [25], the set of virtual feature vectors considered are in the form of the components of class independent GMM (CIGMM). In this comparison, effective number of Gaussian component is empirically chosen as 128. For every component of CIGMM, a feature vector each from the two sets of feature vectors, which has the highest probability of belonging to that component (i.e., value of responsibility term) is selected and a base kernel (such as Gaussian kernel) is computed between the selected feature vectors. In FK [27] GMM with 256 Gaussian component is used for mapping a set of feature vectors onto a Fisher score-space. The Fisher score-space for a class is obtained using the first order derivatives of the log likelihood output of GMM for that class with respect to the GMM parameters. The PSK [28] maps a set of feature vectors onto a high dimensional probabilistic score space. The probabilistic score space for a class is obtained using the posterior probability of components of the GMM built for that class. Effective number of Gaussian component is considered as 256. The accuracies presented in Table II are the best accuracies observed among the Bayes classifier using GMM and SVM-based classifiers with dynamic kernels using different values for their parameters. It is observed that

the SVM-based classifier using the proposed CNN-SLPMK performed better than state-of-the-art results on EMO-DB and FAU-AEC dataset. The proposed approach also has potential to scale up for huge dataset and perform better.

VI. CONCLUSION

In this paper, we proposed a novel CNN architecture with SPP-layer that operate on varying length feature representation of speech signals to perform emotion classification task. The proposed architecture results in varying size set of feature maps. We also proposed a CNN-based segment-level pyramid match kernel (CNN-SLPMK) to match a pair of varying size set of feature maps. The effectiveness of the SVM-based classifier using proposed CNN-SLPMK is demonstrated for speech emotion recognition tasks. The advantage of the proposed approach is that it consider the varying length feature representation of speech signals as input. A limitation of the proposed kernel is that, it requires a CNN model to obtain varying size feature maps. If the pre-trained network or an universal CNN model is available, it reduces the complexity and the resulting kernel will also be very effective. The proposed approaches can be used for classification of varying length speech patterns in the tasks such as spoken language identification, speaker verification, building acoustic models for large vocabulary speech recognition etc.

REFERENCES

- [1] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karay, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [2] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, no. 9–10, pp. 1062–1087, 2011.
- [3] Sadaoki Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.
- [4] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, January 2000.
- [5] A. D. Dileep and C. Chandra Sekhar, “Speaker recognition using pyramid match kernel based support vector machines,” *International Journal for Speech Technology*, vol. 15, no. 3, pp. 365–379, September 2012.
- [6] Shikha Gupta, Veena Thenkanidiyoor, and A. D. Dileep, “Segment-level probabilistic sequence kernel based support vector machines for classification of varying length patterns of speech,” in *Proceedings of International Conference on Neural Information Processing (ICONIP 2016)*. Springer, 2016, pp. 321–328.
- [7] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, “Places: A 10 million image database for scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
- [8] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [9] Krishan Sharma, Shikha Gupta, A. D. Dileep, and Renu Rameshan, “Scene image classification using reduced virtual feature representation in sparse framework,” *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2701–2705, 2018.
- [10] Shikha Gupta, Deepak Kumar Pradhan, Dileep Aroor Dinesh, and Veena Thenkanidiyoor, “Deep spatial pyramid match kernel for scene classification,” in *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM*, 2018, pp. 141–148.
- [11] Tom Sercu and Vaibhava Goel, “Advances in very deep convolutional neural networks for LVCSR,” *arXiv preprint arXiv:1604.01792*, 2016.
- [12] Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu, “Very deep convolutional neural networks for noise robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.
- [13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “ImageNet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [14] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba, “Sun database: Large-scale scene recognition from abbey to zoo,” in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [16] F. Burkhardt, A. Paeschke, M. Rolfs, and W. S. B. Weiss, “A database of German emotional speech,” in *Proceedings of INTERSPEECH*, Lisbon, Portugal, September 2005, pp. 1517–1520.
- [17] S. Steidl, “Automatic classification of emotion-related user states in spontaneous children’s speech,” PhD thesis, Der Technischen Fakultät der Universität Erlangen-Nürnberg, Germany, 2009.
- [18] Donggeun Yoo, Sunggyun Park, Joon-Young Lee, and In So Kweon, “Fisher kernel for deep neural activations,” *arXiv preprint arXiv:1412.1628*, 2014.
- [19] Bin-Bin Gao, Xiu-Shen Wei, Jianxin Wu, and Weiyao Lin, “Deep spatial pyramid: The devil is once again in the details,” *CoRR*, vol. abs/1504.05277, 2015.
- [20] WQ Zheng, JS Yu, and YX Zou, “An experimental study of speech emotion recognition based on deep convolutional neural networks,” in *Proceedings of 2015 International Conference on Affective Computing and Intelligent Interaction (ACII 2015)*. IEEE, 2015, pp. 827–831.
- [21] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran, “Deep convolutional neural networks for LVCSR,” in *Proceeding of 2013 IEEE international conference on Acoustics, speech and signal processing (ICASSP 2013)*. IEEE, 2013, pp. 8614–8618.
- [22] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, 2006, vol. 2, pp. 2169–2178.
- [23] Shikha Gupta, A. D. Dileep, and Veena Thenkanidiyoor, “Segment-level pyramid match kernels for the classification of varying length patterns of speech using svms,” in *Proceedings of 24th European Signal Processing Conference (EUSIPCO 2016)*. IEEE, 2016, pp. 2030–2034.
- [24] Chih-Chung Chang and Chih-Jen Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.
- [25] A. D. Dileep and C. Chandra Sekhar, “GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, Aug 2014.
- [26] Sabri Boughorbel, Jean Philippe Tarel, and Nozha Boujemaa, “The intermediate matching kernel for image local features,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2005)*, Montreal, Canada, July 2005, pp. 889–894.
- [27] Nathan Smith and Mark Gales, “Speech recognition using svms,” in *Advances in neural information processing systems*, 2002, pp. 1197–1204.
- [28] K-A. Lee, C.H. You, H. Li, and T. Kinnunen, “A GMM-based probabilistic sequence kernel for speaker verification,” in *Proceedings of INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 294–297.

Full-Reference Video Quality Assessment Using Deep 3D Convolutional Neural Networks

Sathya Veera Reddy Dendi

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India
ee16resch01003@iith.ac.in*

Gokul Krishnappa

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India
ee16mtech11006@iith.ac.in*

Sumohana S. Channappayya

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India
sumohana@iith.ac.in*

Abstract—We present a novel framework called Deep Video Quality Evaluator (DeepVQUE) for full-reference video quality assessment (FRVQA) using deep 3D convolutional neural networks (3D ConvNets). DeepVQUE is a complementary framework to traditional handcrafted feature based methods in that it uses deep 3D ConvNet models for feature extraction. 3D ConvNets are capable of extracting spatio-temporal features of the video which are vital for video quality assessment (VQA). Most of the existing FRVQA approaches operate on spatial and temporal domains independently followed by pooling, and often ignore the crucial spatio-temporal relationship of intensities in natural videos. In this work, we pay special attention to the contribution of spatio-temporal dependencies in natural videos to quality assessment. Specifically, the proposed approach estimates the spatio-temporal quality of a video with respect to its pristine version by applying commonly used distance measures such as the l_1 or the l_2 norm to the volume-wise pristine and distorted 3D ConvNet features. Spatial quality is estimated using off-the-shelf full-reference image quality assessment (FRIQA) methods. Overall video quality is estimated using support vector regression (SVR) applied to the spatio-temporal and spatial quality estimates. Additionally, we illustrate the ability of the proposed approach to localize distortions in space and time.

Index Terms—3D ConvNets, full reference video quality assessment, human visual system and spatio-temporal features.

I. INTRODUCTION

Robust objective video quality assessment algorithms have an important role to play in the design of a wide variety of algorithms that range from encoder design to denoising to camera pipeline optimization to quality of experience based resource allocation, to name a few. Quality assessment algorithms become important given the phenomenal growth of mobile devices and their multimedia content generating ability, coupled with social media platforms that allow for the sharing of multimedia content. Further, several camera makers (mobile and otherwise) claim the quality of images and videos they generate to be a differentiating factor compared to the competition.

Since humans are ultimate receivers of visual content, it is important to consider the ratings given by the subjects.

This work is supported under Visvesvaraya PhD scheme by the Media Asia Lab, Ministry of Electronics and Information Technology, Government of India, and we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

978-1-5386-9286-8/19/\$31.00 © 2019 IEEE

The average ratings of subjects are measured as mean opinion scores (MOS) and another well-accepted measure is a difference of mean opinion score (DMOS). DMOS is defined as the difference of reference signal MOS to distorted signal MOS. Since subjective ratings are always time consuming and expensive, it is important to have an alternative automatic or objective mechanism which predicts the quality of the video in correlation with the subjective ratings. Objective metrics are broadly classified into three categories based on the availability of reference video. If the metric has access to the full reference signal it is called as the full-reference (FR). If it has to partial information then it is called reduced-reference (RR) and if there is no access to the reference signal is called no-reference (NR) objective metrics. In this paper, we address the problem of full reference video quality assessment (FRVQA) using 3D ConvNets [1]. A brief review of relevant literature of FRVQA techniques follows.

A simple way of addressing the objective VQA problem is applying the objective image quality assessment techniques on a frame-by-frame basis and averaging the frame level quality scores. Such approaches completely ignore motion which is crucial in video perception. So, it is important to consider the motion information while designing the objective VQA techniques. We briefly reviewed the popular and relevant VQA techniques like, The motion-based video integrity evaluator (MOVIE) index [2] is a FRVQA metric that quantifies the deviation in the optical flow plane in the distorted video with respect to the reference video across spatio-temporal frequency bands and pools them to estimate quality. VQM-VFD [3] and FLOSIM [4] measure video quality by quantifying spatial and temporal distortions in a video individually and pooling these scores to arrive at one spatio-temporal score. A common theme running across all these excellent metrics is that they have not only identified (hand picked) very effective spatial and temporal features to discern distortion but have also been able to pool these spatial and temporal scores effectively, albeit based on heuristics in several cases.

The drawback with most of the existing FRVQA solutions is that they typically evaluate spatial and temporal quality separately and employ a pooling strategy to come up with the overall quality score. However, the intensities of natural videos display rich dependencies in space and time that could

be used for constructing a better FRVQA metric.

Recent advances in deep neural networks provide a natural and logical setting for building a quality assessment framework. The ConvNet representations of images and videos are generic, simple, compact and efficient [1] due to the fact that these networks are trained using natural images and videos. ConvNets have been successfully applied to 2D no-reference image quality assessment (NRIQA) [5] [6], 3D stereoscopic NRIQA [7] and reduced-reference video quality assessment (RRVQA) [8]. A neural network based FRIQA algorithm was proposed by Bosse et al. [9]. Given the success of the 2D ConvNets in the IQA problem, we propose a natural extension to the VQA task using 3D ConvNets which can capture the spatio-temporal dependencies of natural videos. In the proposed method, deep 3D ConvNets are used to extract spatio-temporal representations of pristine and test videos and the deviation of test video features relative to pristine video features is used to estimate the perceptual quality of the test video.

The rest of the paper is organized as follows: the proposed framework is presented next, followed by experimental analysis. Results and discussions and conclusions follow subsequently.

II. PROPOSED FRAMEWORK

In this section, we present the proposed framework along with the motivation to design FRVQA algorithm using 3D ConvNets. As discussed earlier, most existing FRVQA methods are designed using handcrafted features for the computation of spatial and temporal quality scores or spatiotemporal quality scores. Since feature selection is “more of an art than science,” determining the perceptual importance of features is not a straightforward task and usually relies on inspiration from the human visual system (HVS) and other heuristics. To address this problem from a different perspective, we adopt a learning based feature extraction strategy. Specifically, we use 3D ConvNets for spatio-temporal feature extraction. The motivation for this approach is detailed next.

A. Motivation

The motivation to use 3D ConvNets for spatio-temporal feature extraction is that they share weights in both space and time that enables them to extract generic and compact representations of natural videos [1]. We also draw an analogy between 3D ConvNets [10] and the HVS based on the hypothesis that the volumetric kernels of 3D ConvNets extract spatiotemporal features and serve as a combination of the V1 and MT areas in the HVS. V1 is region in primary visual cortex is where most of the visual information is processed. MT area is a secondary level visual signal processing stage where motion information of the visual signals is processed. The area MT and V1 are interconnected directly and also via other visual processing areas like V2 and V3.

B. Framework

The proposed framework is shown in Fig. 1. It comprises of two sections - one for spatial quality estimation of video

on a frame-by-frame basis and the other for spatio-temporal quality estimation by taking video volumes as input. For spatial quality estimation we use the robust MS-SSIM index [11], and for spatio-temporal quality estimation, we measure the deviation between pristine and test video 3D ConvNet features. The spatial quality score and spatio-temporal quality scores are pooled using SVR to estimate the overall quality of the distorted video with respective to pristine video.

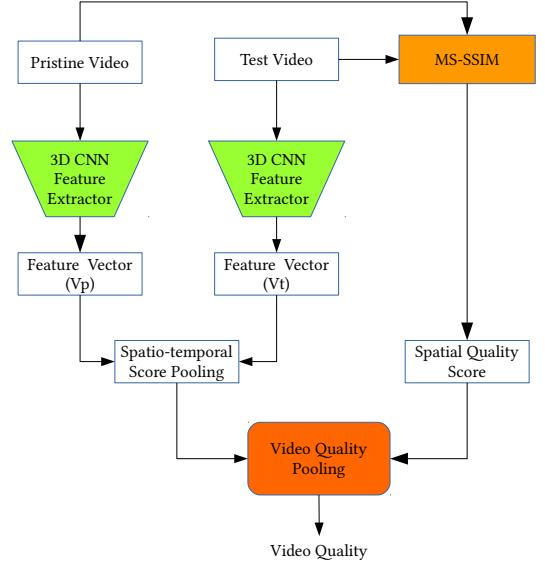


Fig. 1: DeepVQUE framework for FRVQA.

C. Spatial quality estimation

The proposed framework is validated using the MS-SSIM index as the spatial quality metric. Let the spatial quality score of the i^{th} test video frame relative to the reference be denoted by S_i . Then the overall spatial quality of video is defined Q_s as,

$$Q_s = \frac{1}{N} \sum_{i=1}^N S_i, \quad (1)$$

where N is the total number of frames in the video. While we have used the MS-SSIM index to estimate spatial quality, this could easily be replaced by another FRIQA algorithm.

D. Spatio-temporal quality estimation

To estimate the spatio-temporal quality of video, we extract features by feeding forward both pristine and distorted video through pre-trained models. The features of pristine and distorted video are compared using standard distance measures such as the l_1 or the l_2 norm. The proposed framework is flexible and effective due to its ability to adopt the state-of-the-art pre-trained deep ConvNets models as feature extractors and ubiquitous distance measures for quality estimation. In 3D ConvNets, convolution and pooling layers are in the form of

volumes and they perform volumetric convolution and pooling operations which enables the extraction of spatio-temporal features of video volumes.

Let p and d denote the pristine and distorted video respectively that are fed forward through the 3D ConvNet model Z . The feature vectors at the intermediate layers of our interest is denoted by $V_p = [v_{1p}, v_{2p}, \dots, v_{Np}]^T$ and $V_d = [v_{1d}, v_{2d}, \dots, v_{Nd}]^T$ of both pristine and distorted videos respectively. The spatio-temporal quality of the distorted video is estimated using a distance measure $d(V_p, V_d)$. We demonstrate the effectiveness and the flexibility of proposed framework with simple distance measures like mean absolute error (MAE) or the l_1 norm and mean squared error (MSE) or the l_2 norm.

$$\text{MAE} = d(V_p, V_d) = \frac{1}{N} \sum_{i=1}^N |v_{ip} - v_{id}|, \quad (2)$$

$$\text{MSE} = d(V_p, V_d) = \frac{1}{N} \sum_{i=1}^N (v_{ip} - v_{id})^2. \quad (3)$$

In general, 3D ConvNets are designed for a specific resolution of videos. However, videos from validation datasets may not have the same resolution as that accepted by the pre-trained 3D ConvNets. To overcome this issue, we follow Algorithm 1.

Algorithm 1: Spatio-temporal VQA algorithm

- Input:** Pristine video and distorted video
Output: Spatio-temporal quality of distorted video
- 1 Load pre-trained 3D ConvNet model
 - 2 Divide videos into M non-overlapping volumes of size accepted by the pre-trained model.
 - 3 Initialize volume count i to 0.
 - 4 Feed forward: Extract the feature vectors of i^{th} pristine and distorted video volumes, at the intermediate layer of our interest as $V_p^i = [v_{1p}^i, v_{2p}^i, \dots, v_{Np}^i]^T$ and $V_d^i = [v_{1d}^i, v_{2d}^i, \dots, v_{Nd}^i]^T$.
 - 5 Spatio-temporal quality score of video volume is given by $d(V_p^i, V_d^i)$
 - 6 $i = i + 1$. If $i < M$, goto step 5.
 - 7 Overall spatio-temporal quality of video is given by,

$$Q_{st} = \frac{1}{M} \sum_{i=1}^M d(V_p^i, V_d^i). \quad (4)$$

E. Overall quality estimation

The overall quality of video is estimated using support vector regression (SVR). The spatial quality score and spatio-temporal quality score are used to train the SVR against DMOS scores of the VQA datasets. The standard procedure followed in the literature is to split the dataset into 80% training samples and the remaining 20% for testing. A similar procedure is adopted in evaluating the performance of the proposed approach.

III. EXPERIMENTAL ANALYSIS

In this section, we present a detailed analysis of the spatio-temporal feature extraction and its performance. Pretrained 3D ConvNet models are essential to evaluate the performance of the proposed framework. To get a set of pretrained models, we used 3D convolutional autoencoder (CAE) like the neural network architecture where weights are learned by using videos from the KoNViD-1k [12] database in an unsupervised setting. The network architecture details are given in Table I and the features at the bottle neck are used for spatio-temporal quality estimation. The proposed framework is trained and tested with different input video volume sizes and the performance evaluation is given in Table II. From these results, it is clear that the 3D ConvNet features do indeed possess the ability for discerning perceptual quality.

TABLE I: 3D CAE model summary.

3D CAE	Layer	No.of filters	Parameters
Encoder	Convolution3D (3,3,3)	32	896
	MaxPooling3D (2,2,2)	-	0
	Convolution3D (3,3,3)	16	13840
	MaxPooling3D (2,2,2)	-	0
	Convolution3D (3,3,3)	8	3468
	MaxPooling3D (2,2,2)	-	0
Decoder	Convolution3D (3,3,3)	8	1736
	UpSampling3D (2,2,2)	-	0
	Convolution3D (3,3,3)	16	3472
	UpSampling3D (2,2,2)	-	0
	Convolution3D (3,3,3)	32	13856
	UpSampling3D (2,2,2)	-	0
	Convolution3D (3,3,3)	1	865
Total trainable parameters			38129

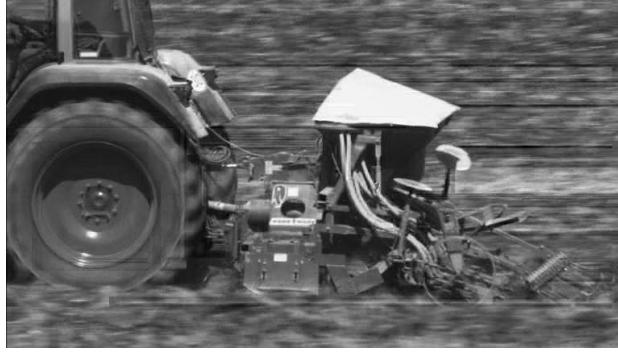
TABLE II: Linear correlation coefficient (LCC) based performance measurement on spatio-temporal quality scores with MAE as distance measure.

Model	Input size (height \times width \times time)	LIVE SD	LIVE Mobile
Model-1	48 \times 48 \times 8	0.81	0.67
Model-2	80 \times 80 \times 16	0.84	0.69
Model-3	120 \times 120 \times 4	0.85	0.72

We also observe that the right set of network parameters can further improve the performance of the proposed framework. To support the above claim, we also evaluated the proposed framework using the state-of-the-art pretrained 3D ConvNet called C3D [1].

A. Localization of spatio-temporal quality of video

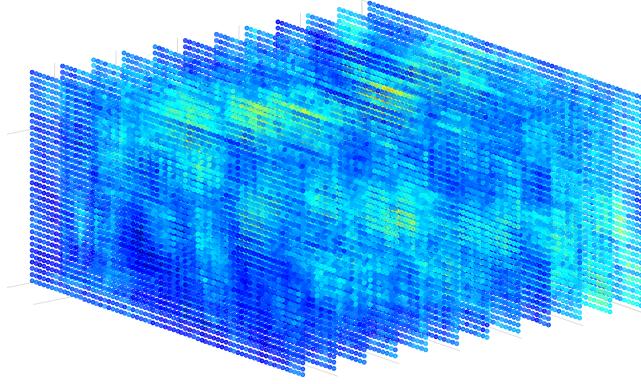
Since the proposed approach works on the spatio-temporal video volumes, it gives the flexibility to localize distortions in space-time, in particular, the (x, y, t) location where x, y are spatial coordinates and t is the time coordinate. Fig. 2 shows the localization of the distortions in space-time of a sample high and low quality video. The distortion maps in Fig. 2 are computed by using pretrained Model-2, details about Model-2 are given in Table II. Distortion intensity at particular



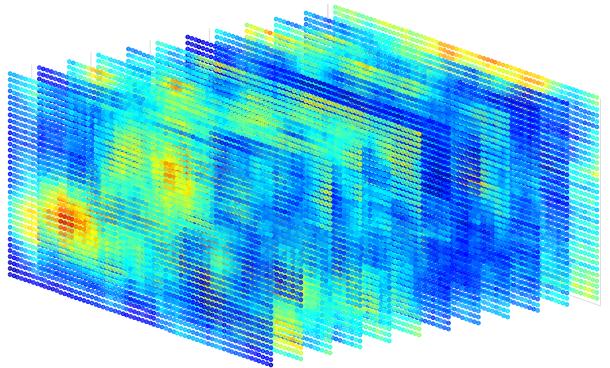
(a) High quality video.



(b) Low quality video.



(c) Distortion map of high quality video.



(d) Distortion map of low quality video.

Fig. 2: Localization of spatio-temporal quality of videos (best view with zoom and color display).

pixel location is estimated using the spatiotemporal features difference between the distorted video and corresponding pristine video with 70-pixel overlap in spatial and no overlap in the temporal direction. For illustration, we have taken two videos from LIVE SD [13] VQA dataset at two extreme levels of perceptual quality (high quality and low quality).

B. Using C3D network for VQA

The proposed framework is also validated using the 3D pre-trained ConvNet model. Unlike the variety of deep 2D ConvNets available for images, deep 3D ConvNets are only recently emerging. To the best of our knowledge, the only freely available pre-trained model is C3D [1] to validate the proposed framework. C3D accepts video volume as input at a resolution of $171 \times 128 \times 16$ and has been trained over 1 million YouTube sports videos which include natural scenes. We evaluated the performance of the proposed framework on videos by extracting features at first fully connected layer of the C3D model. The performance of the DeepVQUE framework using C3D is compared in on our performance analysis.

IV. RESULTS AND DISCUSSION

We report the performance of the proposed DeepVQUE on VQA datasets LIVE-SD [13], EPLE PoliMI [14] and LIVE Mobile [15] using linear correlation coefficient (LCC) and

Spearman rank order correlation coefficient (SRCC) between subjective scores and objective scores obtained by the proposed approach. The details of the above VQA datasets are given in Table III. Table IV and V show the performance of the proposed FRVQA framework on SD and HD datasets respectively by using pretrained models like Model-2 and C3D with l_1 and l_2 as distance measures and MS-SSIM as spatial quality estimator. The Fig. 2 shows the effectiveness of the proposed approach in localizing the distortions of the videos in space and time.

Through these performance numbers, it is clear that the proposed approach based on using 3D ConvNet features is indeed promising. We believe that the performance can be improved by fine-tuning the weights on a larger set of data points. Further, given that the FRVQA algorithm extracts features in the feedforward mode, it could be implemented very efficiently using parallelization on GPUs. Lastly, it shows a way forward that does not need the significant efforts involved in identifying handcrafted features.

V. CONCLUSION

DeepVQUE is an effective complementary framework to existing handcrafted feature based FRVQA methods. The proposed framework is robust and flexible since the choice of the deep 3D ConvNet for feature extraction and the distance measure for quality estimation can easily be changed/tuned. The

TABLE III: Details of existing video quality assessment datasets.

Dataset	No.of. videos	Resolution	Frame rate	Type of distortions	Lighting conditions
LIVE SD [13]	150	768 × 432	25/50 fps	MPEG-2, H.264, error prone IP and wireless networks	Day light
EPFL Polimi [14]	144	CIF resolution(352 × 288) and 4CIF resolution (704 × 576)	30 fps	Transmission errors	Day light
LIVE Mobile [15]	160	1280 × 720	30 fps	Compression, wireless packet-loss, rate-adapted and temporal dynamics	Day light

TABLE IV: Performance on the LIVE-SD [13] and EPFL PoliMI [14] datasets.

Method	LIVE-SD		PoliMI	
	LCC	SRCC	LCC	SRCC
MOVIE [2]	0.81	0.79	0.93	0.92
VQM [16]	0.72	0.70	0.84	0.83
FLOSIM (BA) [4]	0.86	0.85	0.95	0.96
DeepVQUE: Model-2 + l_1	0.82	0.79	0.91	0.91
DeepVQUE: Model-2 + l_2	0.82	0.80	0.89	0.89
DeepVQUE: C3D + l_1	0.85	0.69	0.93	0.92
DeepVQUE: C3D + l_2	0.85	0.70	0.91	0.89

TABLE V: Performance on the LIVE Mobile [15] dataset.

Method	Mobile		Tablet	
	LCC	SRCC	LCC	SRCC
MOVIE [2]	0.72	0.64	0.64	0.78
VQM [16]	0.69	0.69	0.58	0.55
FLOSIM (BA) [4]	0.89	0.87	0.75	0.80
DeepVQUE: Model-2 + l_1	0.81	0.79	0.70	0.66
DeepVQUE: Model-2 + l_2	0.83	0.79	0.77	0.71
DeepVQUE: C3D + l_1	0.83	0.80	0.70	0.66
DeepVQUE: C3D + l_2	0.82	0.80	0.70	0.65

competitive performance of DeepVQUE justifies the choice of spatio-temporal features for the VQA task. We believe that the DeepVQUE framework serves as a template for developing future FRVQA algorithms using spatio-temporal representations and helps in localizing distortions in space-time. We also want to make a observation that the existing FRVQA techniques are computationally expensive, but the proposed approach involves simple feedforward operation during the testing phase, so this can be used in real time applications. Proposed approach also allows to deploy the state-of-the-art 3D ConvNet models to improve the performance further.

REFERENCES

- [1] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- [2] Kalpana Seshadrinathan and Alan Conrad Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2010.
- [3] Margaret H Pinson, Lark Kwon Choi, and Alan Conrad Bovik, “Temporal video quality model accounting for variable frame delay distortions,” *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 637–649, 2014.
- [4] Manasa K. and Sumohana S. Channappayya, “An optical flow-based full reference video quality assessment algorithm,” *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, June 2016.
- [5] Le Kang, Peng Ye, Yi Li, and David Doermann, “Convolutional neural networks for no-reference image quality assessment,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [6] Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini, “On the use of deep learning for blind image quality assessment,” *CoRR*, vol. abs/1602.05531, 2016.
- [7] Wei Zhang, Chenfei Qu, Lin Ma, Jingwei Guan, and Rui Huang, “Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network,” *Pattern Recognition*, vol. 59, pp. 176 – 187, 2016, Compositional Models and Structured Learning for Visual Recognition.
- [8] Patrick Le Callet, Christian Viard-Gaudin, and Dominique Barba, “A convolutional neural network approach for objective video quality assessment,” *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1316–1327, Sept 2006.
- [9] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, “Full-reference image quality assessment using neural networks,” in *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*, short paper, 2016.
- [10] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, “3d convolutional neural networks for human action recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.
- [11] Z Wang, E P Simoncelli, and A C Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc 37th Asilomar Conf on Signals, Systems and Computers*, Pacific Grove, CA, Nov 9-12 2003, vol. 2, pp. 1398–1402, IEEE Computer Society.
- [12] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe, “The konstanz natural video database (konvid-1k),” in *Quality of Multimedia Experience (QoMEX)*, 2017 Ninth International Conference on. IEEE, 2017, pp. 1–6.
- [13] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan C. Bovik, and Lawrence K. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.
- [14] Francesca De Simone, Marco Tagliasacchi, Matteo Naccari, Stefano Tubaro, and Touradj Ebrahimi, “A h. 264/avc video database for the evaluation of quality metrics,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2430–2433.
- [15] Anush Krishna Moorthy, Lark Kwon Choi, Alan Conrad Bovik, and Gustavo De Veciana, “Video quality assessment on mobile devices: Subjective, behavioral and objective studies,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 652–671, 2012.
- [16] Margaret H Pinson and Stephen Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on broadcast ing*, vol. 50, no. 3, pp. 312–322, 2004.

Deep Learning-Based Modulation Classification Using Time and Stockwell Domain Channeling

Shrishail M Hiremath, Sambit Behura

Department of Electronics and Communication Engineering
National Institute of Technology Rourkela
Rourkela, India
Email: hiremaths@nitrkl.ac.in,
sambitbehura001@gmail.com

Siddharth Deshmukh

Department of Electronics and Communication Engineering
National Institute of Technology Rourkela
Rourkela, India
Email: deshmukhs@nitrkl.ac.in

Subham Kedia

Department of Computer Science and Engineering
National Institute of Technology Rourkela
Rourkela, India
Email: subhamkedia799@gmail.com

Sarat Kumar Patra

Indian Institute of Information Technology Vadodara
Vadodara, India
Email: skpatra@iiitvadodara.ac.in

Abstract— Deep learning techniques have recently exhibited unprecedented success in classification problems with ill-defined mathematical models. In this paper, we apply deep learning for RF data analysis and classification. We present a novel method of using I-Q time samples to form images with ‘Time and Discrete Orthonormal Stockwell Transform Domain Channels’ which are used for training a convolutional neural network (CNN) for radio modulation classification. Also, a concept inspired from transfer learning is used in extending the number of output classes of the CNN, which helps the network to estimate the approximate SNR of the input signal as well and further improve the classification accuracy. Such a network trained on Time and Stockwell Channeled Images performs superior to similar networks that are trained on just raw I-Q time series samples or time-frequency images, especially when training samples are less. The network achieved an overall classification accuracy of 97.3% at 8 dB SNR over a class of 10 radio modulation schemes (for both digital and analog systems). The study shows that such a trained network can be well applied to achieve high classification accuracies at low and moderate SNR scenarios.

Keywords—deep learning, convolutional neural network, modulation classification, Stockwell transform, discrete orthogonal Stockwell transform, time and Stockwell domain channeling.

I. INTRODUCTION

RADIO data available from an antenna is often easily captured, but in the modern day it is difficult to label and curate the data accurately from the complex high-data rate RF information. The strategies adopted for such tasks are often time-consuming, and their implementations are not precise under varying environmental conditions. Hence, blind radio signal recognition and identification at the receiver end has turned out to be a very useful and important tool in dense, multi-user scenarios. Fast labeling and understanding of the radio spectrum can provide added advantages like optimized spectrum utilization, minimized and identifiable interference, spectrum policy enforcement, and implementing efficient spectrum sensing and coordination systems. Hence it has enabled radio fault detection, spectrum interference monitoring, dynamic spectrum access, opportunistic mesh networking and numerous other fields in communication systems.

Modulation classification is the process of blindly identifying and differentiating radio signals at the receiver end, as a step towards understanding the type of communication schemes being used by the transmitters in the vicinity. Modulation recognition or classification is a front-end tool in number of applications like link adaptation, modern military signal intelligence systems, spectrum monitoring systems, unmanned aerial drones, dynamic spectrum access, cognitive radio, and cellular standards like LTE-Advanced.

The last two decades have seen wide variety research on developing novel methods and algorithms for automatic radio modulation classification/recognition. Many of these are carefully designed feature extraction based techniques which project the received signal on a low-dimensional feature space in which compact decision boundaries can help differentiate one radio modulation from the other [1]. Modulation classification can also be performed using generative algorithms based on probabilistic models like Naïve Bayes [2], hidden Markov models obtained with maximum likelihood estimation [3] and methods that uses likelihood ratio tests such as the average likelihood ratio test (ALRT) [4], generalized likelihood ratio test (GLRT) [5] and the hybrid likelihood ratio test (HLRT) [6]. Integrated cyclic moment-based features [7] and features based on CSS (Concatenated Sorted Symbols) [8] are widely popular for forming analytically derived decision trees to sort modulations into different classes. Despite its robustness against noise and interference, cyclostationary analysis of a signal has a high computational cost and is not efficient for quick labeling and real-time modulation classification [7].

In the past few years, there have been massive improvements and developments in neural network architectures and optimization algorithms. Deep neural networks have pushed performance boundaries of machine learning tasks in a variety of applications. This deep learning trend, which is quite popular in computer vision or text processing, is yet to be adequately explored and fully applied to complex temporal radio signal datasets. Moreover, in case of RF data, the type of samples that the network is to be trained on, and whether or not some kind of pre-processing on the time samples might improve

training performance needs to be investigated as well. In this work we propose a novel method of using I-Q time samples to form images with ‘Time and Discrete Orthonormal Stockwell Transform Domain Channels’ which are then used for training a convolutional neural network (CNN) for the task of radio modulation classification. This type of training on the network proves to be really efficient, especially when the number of training samples is less.

The organization of this paper is as follow: Section II mentions some notable papers that have used deep neural networks for modulation classification, and also gives a brief background on the need for time-frequency analysis and the Stockwell transform. Section III presents an overview of the available radio machine learning datasets. Section IV presents the details about the proposed classification approach i.e. system model, Time & Stockwell Domain Channeling, the CNN architecture and the concept of Extended Output Classes. Section V presents the experimental results and their detailed analysis. Section VI concludes this paper.

II. BACKGROUND

A. Deep Architectures for RF Data Classification

Applying deep learning to a problem like modulation classification involves selecting a network architecture and hyper-parameters, training the network to optimize weights that minimize loss, and applying the trained network to the problem at hand. [9] presents a survey of the various deep learning architectures inspired from computer vision and natural language processing that can be applied to the task of modulation classification. Some deep architectures that have found success in radio signal identification include Convolutional and Residual Networks [10] [11], Recurrent and LSTM networks [12] and Heterogeneous Deep Fusion Models [13].

B. Need for Time-Frequency Analysis – The Stockwell Transform

One major drawback of Fourier transform (FT) as a spectral analysis tool is that it produces only the time-averaged spectrum, thus making it unfavorable for applications where local information is preferred (e.g., signal de-noising, compression, phase analysis) [14]. Thus in recent years, more advanced representations known as joint time-frequency representations have been adopted [15].

The wavelet transform [16] is a time-frequency decomposition that applies local decomposition filters to a signal on multiple scales. But, even though the term “scale” can be approximately interpreted as “frequency,” there is no way to extract proper frequency information from the scale information [14].

The Stockwell transform (ST, also popularly known as the S-transform) [17] [18] [19] [15] is a time-frequency decomposition that provides absolutely referenced phase information. Here, the summation of the coefficients for a fixed frequency gives the exact Fourier coefficient for that frequency [14].

Consider a one dimensional signal $x(t)$. The ST of $x(t)$ is defined as the FT of the product between $x(t)$ and a Gaussian window function.

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t) \frac{|f|}{\sqrt{2\pi}} e^{-\frac{(\tau-t)^2 f^2}{2}} e^{-j2\pi ft} dt \quad (1)$$

By properties of the Gaussian function, the relationship between $S(\tau, f)$ and $X(f)$ (FT of $x(t)$) is given as

$$\int_{-\infty}^{\infty} S(\tau, f) d\tau = X(f) \quad (2)$$

Hence, the summation of Stockwell coefficients along the time axis gives the FT of the signal. The original signal $x(t)$ can be recovered by calculating the inverse FT of $X(f)$ [14].

However, it is known that the Stockwell transform (also discrete ST) is highly redundant and thus it needs a large amount of time and storage space even for a moderately long signal. For example, a signal of length N , generates N^2 coefficients through the discrete ST. As a solution to reduce this redundancy, the time-frequency domain can be partitioned into N regions, and each region can be represented by one coefficient [14]. This is the strategy adopted by the discrete orthonormal Stockwell transform (DOST) [20], thus making its computation simpler. The DOST coefficients can be computed by taking the vector dot-product of the input signal with a set of N basis vectors, which gives it a computational complexity $O(N^2)$ [14]. Let a region in the time-frequency domain be described by a set of parameters: v specifies the center of each frequency band, β is the width of the band and τ specifies the point in time. Using these parameters the k^{th} basis vector is defined as

$$D[k]_{[v, \beta, \tau]} = \frac{1}{\sqrt{\beta}} \sum_{f=v-\frac{\beta}{2}}^{v+\frac{\beta}{2}-1} e^{-j2\pi \frac{k}{N} f} e^{j2\pi \frac{\tau}{\beta} f} e^{-j\pi} \quad (3)$$

For $k = 0, \dots, N-1$. An algorithm to compute DOST through a fast method is presented in [14].

III. RADIO MACHINE LEARNING DATASETS

All datasets used in this work are provided by DeepSig Inc., and are licensed under the Creative Commons Attribution – NonCommercial – ShareALike 4.0 License (CC BY-NC-SA 4.0). DeepSig has created some standard datasets which can be used by scientists and engineers for original and reproducible research. These datasets give scope to machine learning researchers to dive directly into new and important technical areas in radio signal processing without the need for collecting or generating new datasets [21].

Dataset RadioML 2018.01A by DeepSig Inc. includes both synthetic simulated channel effects and over-the-air recordings of 24 digital and analog modulation types which has been heavily validated. This dataset was used in [11] which provides additional details and description of the dataset. Data are stored in hdf5 format as complex floating point values, with 2 million examples, each 1024 samples long. The included modulation classes are 32PSK, 16APSK, 32QAM, FM, GMSK, 32APSK, OQPSK, 8ASK, BPSK, 8PSK, AM-SSB-SC, 4ASK, 16PSK, 64APSK, 128QAM, 128APSK, AM-DSB-SC, AM-SSB-WC, 64QAM, QPSK, 256QAM, AM-DSB-WC, OOK and 16QAM [21].

IV. PROPOSED CLASSIFICATION APPROACH

A. System Model

Fig. 1 presents the complete system model that is being used for our classification approach. A RF I-Q image is a matrix containing fixed number of samples (1024 in this case) of the in-phase and quadrature-phase components of the received signal, arranged into rows. This 2×1024 matrix/image serves as the input to the system model. The DOST block performs row-wise DOST (separately for I-samples and Q-samples) on its input image and then takes the absolute values of each transformed complex values. The output of this block is another 2×1024 image which contains Time-Frequency Domain information about the I and Q samples. The input and output images of the DOST block are then fed into the Time and DOST Data Channeling block which forms $2 \times 1024 \times 2$ images by assigning the time samples and DOST processed samples to different channels of its output image. This final image is then given as the input to the first layer of the trained convolutional neural network to perform the classification.

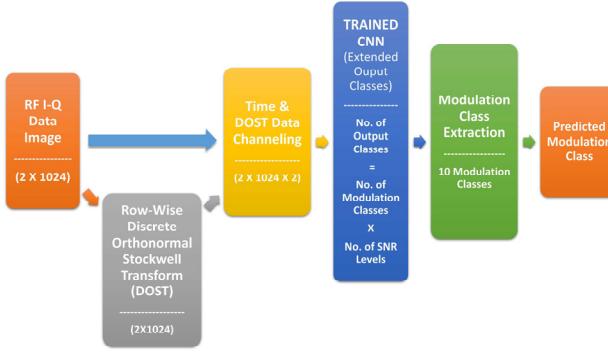


Fig. 1. Proposed System Model of Using Time and DOST Data Channeled Images and Extended Output Classes for CNN-based Radio Modulation Classification.

Before training the CNN, the entire pre-processing is done for all the training and validation I-Q time images as well, to convert them into their equivalent time and DOST channelled images. A detailed explanation of the architecture is presented in sub-section C of this section. The output layer of the CNN predicts one among the extended classes which are labelled according to both modulations as well as the SNR levels considered. This prediction of the extended class is then further processed in the Modulation Class Extraction block to extract the final modulation label for the input test data.

B. Time and Discrete Orthonormal Stockwell Transform Domain Channeling for RF I-Q Images

As described in the system model in sub-section A of this section, a new kind of pre-processing is incorporated in our approach to make the deep convolutional network learn signal features more prominently and efficiently. The output image of the DOST block contains all the time-frequency information from which the network can learn more features.

The physical intuition behind such pre-processing is that missing out on either the time data or time-frequency data could lead to loss of important features that the network could have learnt from a union of information of both the domains provided in a compact form. This led to the idea of creating two different channels in the same

image, similar to RGB channels in a digital photograph. One channel holds the data from time domain I-Q samples and the other channel holds the data from the DOST processed time-frequency domain image. This gives the final image a depth of 2, and the output size is $2 \times 1024 \times 2$. Moreover when the different kernels in a convolutional layer of the CNN train over such images, some of them might adapt to learn time features while some might adapt to learn time-frequency features. This was the inspiration behind the core idea of channeling multiple domain data. An example of the Time and DOST Domain Channeling for 16 I-Q time samples is shown below in Fig. 2.

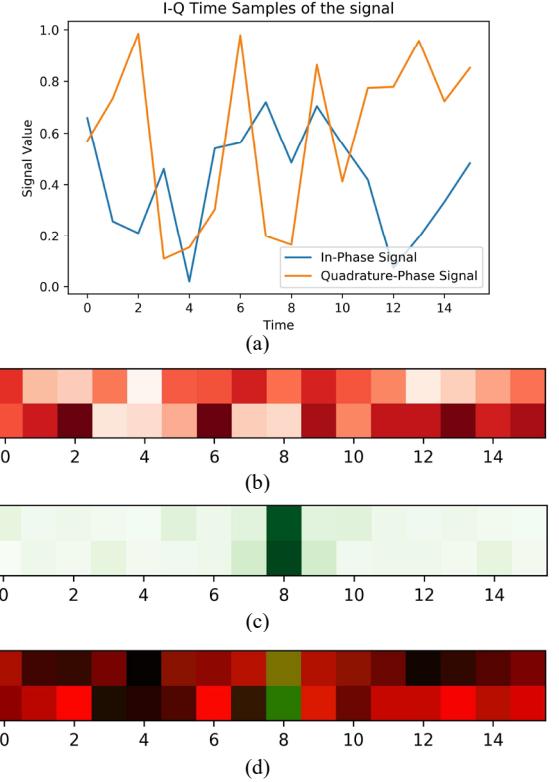


Fig. 2. (a) Plots of In-phase and Quadrature-phase time samples of the signal. (b) Visual representation of the RF I-Q image as the Red-Channel. (c) Visual representation of output of the DOST block as the Green-Channel. (d) Visual representation of the Time and DOST Domain Channeled Image in the form of an RGB image.

(*Note: White represents zero value in (b) and (c), but black represents zero value for RGB in (d) i.e. White in (a) + White in (b) = Black in (d))

Fig. 2(d) is just an equivalent representation of the Time-DOST Channeled image shown in a RGB format, and all the values in the blue channel are assumed zeros.

C. The CNN Architecture and Extended Output Classes

An eight layer convolutional neural network is proposed for the classification task in this work. It consists of 5 convolutional layers and 3 fully connected dense layers (including the output layer). The input image of size $2 \times 1024 \times 2$ is fed to the first convolutional layer (Conv) which has 128 filters, each of size 2×5 . The activation function used is ReLU, and appropriate zero padding is done to keep the output of the first layer the same size as that of the input image. The second, third and fourth convolutional layers are identical to the first layer. The fifth convolutional layer is different from the other four only with respect to the filter size i.e. 2×7 . The sixth and the seventh layers are fully connected dense layers (FC Dense)

with 256 neurons each and activation function used is ReLU. The eighth layer is the output layer, a fully connected dense layer with number of neurons equal to number of output classes, and SoftMax is used as the activation function. Here 10 modulations are considered for classification with the SNR levels varying from -8 dB to +8 dB in increments of 2 dB. Hence the number of extended output classes is 90 (No. of modulation classes multiplied by the No. of SNR levels).

Average pooling of pool size 1x4 is performed after the 1st and 2nd convolutional layer, and a pool size of 1x2 is used after the 3rd and 4th layers respectively. No pooling is performed after the 5th convolutional layer. The CNN layout is presented in Table I. Total number of trainable parameters is 1,861,850.

TABLE I. CNN ARCHITECTURE LAYOUT

Layers	Output Shape	Parameters
Input	2 x 1024 x 2	-
Conv 1 (128x2x5), ReLU	2 x 1024 x 128	2,688
Average Pooling (1x4)	2 x 256 x 128	-
Conv 2 (128x2x5), ReLU	2 x 256 x 128	163,968
Average Pooling (1x4)	2 x 64 x 128	-
Conv 3 (128x2x5), ReLU	2 x 64 x 128	163,968
Average Pooling (1x2)	2 x 32 x 128	-
Conv 4 (128x2x5), ReLU	2 x 32 x 128	163,968
Average Pooling (1x2)	2 x 16 x 128	-
Conv 5 (128x2x7), ReLU	2 x 16 x 128	229,504
FC Dense 6 (256), ReLU	256	1,048,832
FC Dense 7 (256), ReLU	256	65,792
FC Dense 8 (90), SoftMax	90	23,130

The filters used in the convolutional layers are of sizes 2x5 and 2x7. The main motive behind using 2-D filters is to allow the kernels to adapt to I and Q data separately.

The dataset used in this approach is a part of the RadioML 2018.01A real world, over the air captured dataset provided by DeepSig Inc. [21]. 10 primary modulations have been extracted from amongst the 24 modulation dataset for received SNR values ranging from -8 dB to 8dB with increments of 2 dB. These modulation classes include BPSK, QPSK, 8PSK, GMSK, 16APSK, 64 APSK, 16QAM, 64QAM, AM-DSB-WC and FM. We shall refer to this dataset as ‘Master Dataset’ in the rest of this article. The reason for choosing these specific modulation schemes is their extent of application in modern day communication systems like broadcast radios, satellite communication, satellite television, WLAN standards, Wi-MAX standards and cellular standards.

The concept of ‘Extended Output Classes’ is inspired from transfer learning. Each sample in the considered dataset is labelled with two tags i.e. the modulation tag and received SNR level tag. The general approach of using a CNN or any other deep neural network architecture for modulation classification on such a dataset has been classifying the input sample according to the modulation classes only. This has been presented in works [10] [11] and [9]. But in this proposed ‘Extended Output Classes’ method, the CNN is trained to predict both the modulation tag as well as the SNR tag of the input sample. This is done

by defining output classes with [Modulation, SNR] labels rather than just [Modulation] labels. For example, if the number of modulation classes is ‘M’ and the number of SNR levels considered in the dataset is ‘N’, then the number of extended output classes would be a product of M and N. The number of classes is increased by a factor N as compared to the general approaches.

In this work, data samples of 10 modulations over 9 SNR levels (-8:2:8 dB) are considered, hence instead of 10 output classes/modulation labels, we have 90 extended output classes which are the [Modulations, SNR] labels. Hence the last fully connected dense layer has 90 neurons, as presented in Table I. The main idea behind classifying on an extended class size is to make the network understand signal features at different SNR levels in a more adaptable manner and to prepare it for the varying SNR scenarios that it might face during testing on an unknown sample. For this the network should first learn to recognize the approximate SNR scenario from the input sample, and then adapt itself accordingly for achieving a better overall classification accuracy. An example of this process is illustrated in Fig. 3, which shows the extended [Modulation, SNR] output classes for a single modulation class ‘BPSK’.

Finally, as the last part of the system model, a modulation extraction block is used to extract only the [Modulation] label from the predicted [Modulation, SNR] labels by the CNN. The output of modulation extraction is one among the 10 classes of modulations that have been considered. This block can be implemented using a simple many-to-one mapping function.

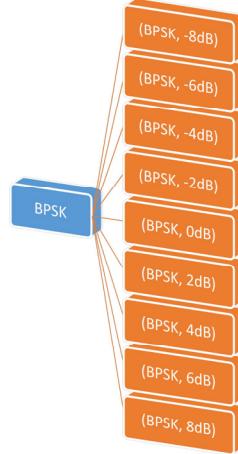


Fig. 3. Extended Output [Modulation, SNR] Classes shown for a single modulation class ‘BPSK’.

V. RESULTS AND ANALYSIS

The model for the CNN architecture described in subsection C of section IV was first built in using Keras. It was then trained, validated and tested on the Master Dataset. It contains a total of 368,640 samples, each sample being a 2x1024 RF I-Q Image in our considered dataset. 85% of the data samples are considered for the training and validation set, i.e. 313,344 samples, out of which 250,675 samples belong to the training set, and 62,669 samples belong to the validation set. The rest 55,296 samples are used as the test set. Training is performed using a categorical cross entropy cost function and an Adam optimizer.

We implement the training and prediction of our network in Keras [22] running on top of TensorFlow on a NVIDIA Cuda powered TESLA V100 16GB GPU in a Google Cloud Compute Engine Virtual Machine (VM) Instance. The VM instance was powered by a quad-core Intel Skylake based processor and 32GB of RAM.

The network was trained and evaluated for four different cases to test the effectiveness of Time & DOST Domain Channeling (T-D-D-C) and Extended Output Classes (E-O-C). Henceforward these two abbreviations shall be used in this article. In the first case, neither T-D-D-C was performed on the data samples, nor were E-O-C used. Hence the neuron count in the final FC dense layer falls to 10 from the previous count of 90. In the second case, T-D-D-C was not performed on the data samples, but E-O-C were used. In the third case, T-D-D-C was performed on the samples, but E-O-C were not used. Hence the network, in this case, is similar to that of the first case, with 10 neurons in the final FC dense layer. In the fourth and the final deciding case, T-D-D-C was performed on that data samples, as well as E-O-C were used.

The classification accuracies achieved by the network for all the four cases, over all the 10 different modulations and for different values of received SNR levels are plotted in Fig. 4. The average overall classification accuracies achieved by the network for all the four cases, over all the 10 different modulations and all values of received SNR levels are shown in Fig. 5.

In the first case, which can be considered to be the baseline case, the network achieved a maximum of 80.49% classification accuracy at 8 dB SNR and overall accuracy of 64.47%. This can be considered to be a very average classification performance.

In the second case the network achieved a maximum of 90.41% classification accuracy at 8 dB SNR and an overall accuracy of 71.82%. This can be considered to be a good push to the CNN performance. As it can be observed from Fig. 4, in this case the accuracies for SNR levels greater than and equal to 0 dB have been boosted, while those for SNR levels below 0 dB have remained more or less similar. Thus it can be inferred that by using extension of output classes, the network seems to have learnt to distinguish between good and bad SNR scenarios by just observing the data samples, and hence the training process has taken place accordingly to optimize performance.

In the third case the network achieved a maximum of 88.07% classification accuracy at 8 dB SNR and overall accuracy of 73.89%. As it can be observed from Fig. 4, in this case, although the maximum achieved accuracy is less as compared to that of the second case, the overall accuracy is more. Moreover the classification accuracies at all the SNR levels seem to have been boosted as compared to that of the first case. Also for SNR levels less than 2 dB, the classification accuracies are more as compared to that of the second case. Hence it can be inferred that by using time & DOST domain channeling on the data samples, the CNN was able to extract and learn more features from both the time domain samples as well as the time-frequency domain samples which accentuated its overall performance.

As a comparison with related papers, the results achieved in the second and third case in this paper already seems to outperform some particular results presented in

[11]. It can be observed by comparing our results with Fig. 7 and Fig. 8 of [11], that for almost the same number of training samples i.e. 250,675 in our case and 240,000 in [11], 10 modulations classes in our case as compared to 11 classes in [11], and 1024 I-Q time samples in a single RF I-Q image in both works, our second and third case already perform much better with respect to classification accuracies as well as understanding SNR scenarios. Also, the maximum SNR considered here is just 8 dB as compared to 20 dB in [11]. This comparison is done while our best case i.e. the fourth case/ proposed method is yet to be analyzed. This is done with an intention to show the novelty and effectiveness of E-O-C and T-D-D-C as independent methods to boost modulation classification performance.

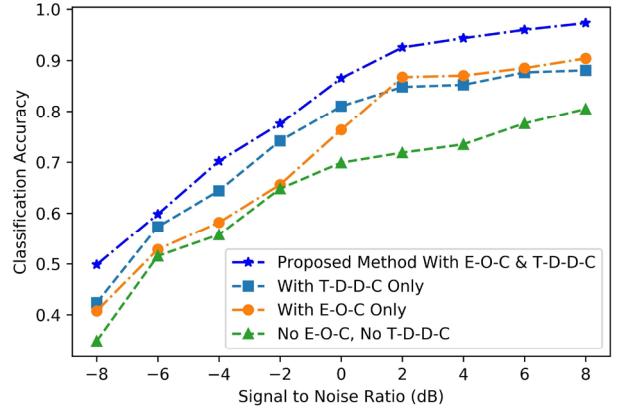


Fig. 4. Classification Accuracy vs. SNR for all the four cases of evaluation.

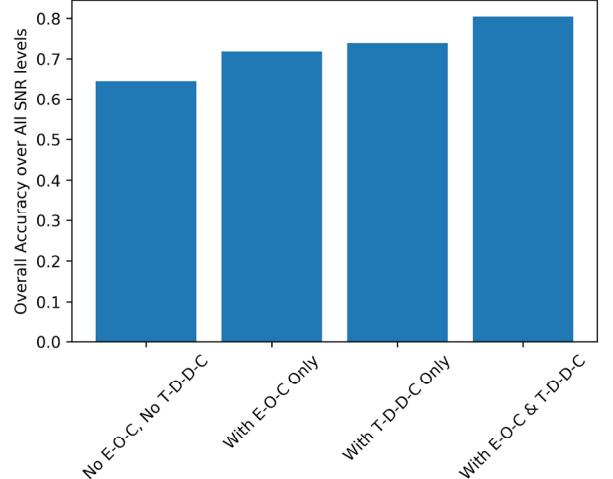


Fig. 5. Overall Classification Accuracy for all the four cases of evaluation.

In the fourth case the network achieved a maximum of 97.30% classification accuracy at 8 dB SNR and an overall accuracy of 80.45%. As it can be observed from Fig. 4, the performance of the CNN, in this case, is far superior to that of all the other cases considered. T-D-D-C and E-O-C combine together to form a strong and effective method to make the CNN learn more time and time-frequency features strongly as well as help it to understand the quality of signals and adapt the optimization process according to varying SNR scenarios. The CNN training process took 84 epochs, each lasting 45 seconds, with a batch size of 512 samples. The validation loss was monitored during the training.

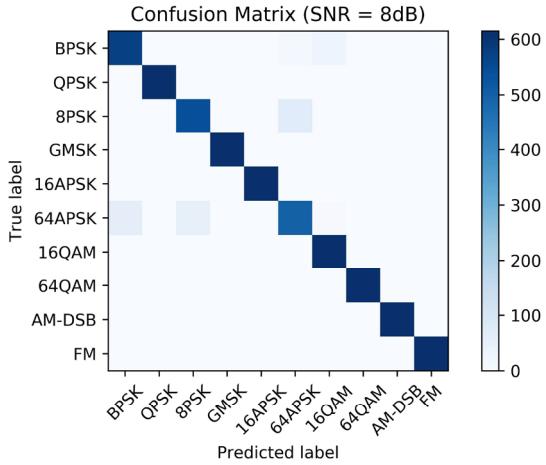


Fig. 6. Confusion Matrix for 10 modulation classes at 8 dB SNR.

To analyze the classification performance with respect to each individual modulation class, the confusion matrix at 8 dB SNR is shown in Fig. 6. It is observed from the figure that a nearly clean diagonal matrix is obtained at 8 dB SNR. The network is able to identify all modulation classes separately with very high accuracy. Though there is some slight confusion between the classes 8PSK and 64APSK, this might be because some of the constellation points are common to both the modulation schemes.

VI. CONCLUSION AND FUTURE SCOPE

Deep learning has seen a lot of development since the last decade. It has had unprecedented success in field like image classification, object recognition, natural language processing, unmanned vehicles, data analytics, and artificial intelligence. But its application in communication systems and devices is yet to be fully explored. This paper presents a way to use CNNs for the task of modulation recognition and also suggests a novel pre-processing method to improve upon their performance. This work might find scope in dynamic spectrum access and spectrum monitoring applications. As a continuation of this work in the future, other deep architectures like RNNs, LSTMs, ResNet structures for CNNs and fusion models as well as concepts of unsupervised learning can be explored in the domain of radio signal classification and identification. Also ensemble models trained individually for encountering samples at different SNR values can be explored to improve classification accuracy at low SNR scenarios. New pre-processing methods like T-D-D-C and suitable changes in the network architecture like E-O-C proposed in this paper can be further explored to accentuate performances of deep networks in specific scenarios. Hardware implementation solutions like FPGA acceleration of CNN as well as acceleration on dedicated low power edge computing devices can also be explored.

REFERENCES

- [1] O. A. Dobre, A. Abdi, Y. Bar-Ness and W. Su, "Survey of automatic modulation classification techniques: classical approaches and new trends," *IET communications*, vol. 1, pp. 137-156, 2007.
- [2] P.-N. Tan and others, *Introduction to data mining*, Pearson Education India, 2007.
- [3] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, vol. 31, Springer Science & Business Media, 2013.
- [4] W. Su, J. L. Xu and M. Zhou, "Real-time modulation classification based on maximum likelihood," *IEEE Communications Letters*, vol. 12, 2008.
- [5] J. L. Xu, W. Su and M. Zhou, "Likelihood-ratio approaches to automatic modulation classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 41, pp. 455-469, 2011.
- [6] P. Panagiotou, A. Anastopoulos and A. Polydoros, "Likelihood ratio tests for modulation classification," in *MILCOM 2000. 21st Century Military Communications Conference Proceedings*, 2000.
- [7] W. A. Gardner and C. M. Spooner, "Signal interception: performance advantages of cyclic-feature detectors," *IEEE Transactions on Communications*, vol. 40, pp. 149-159, 1992.
- [8] F. C. B. F. Muller, C. Cardoso Jr and A. Klautau, "A front end for discriminative learning in automatic modulation classification," *IEEE Communications Letters*, vol. 15, pp. 443-445, 2011.
- [9] N. E. West and T. O'Shea, "Deep architectures for modulation recognition," in *Dynamic Spectrum Access Networks (DySPAN), 2017 IEEE International Symposium on*, 2017.
- [10] T. J. O'Shea, J. Corgan and T. C. Clancy, "Convolutional radio modulation recognition networks," in *International conference on engineering applications of neural networks*, 2016.
- [11] T. J. O'Shea, T. Roy and T. C. Clancy, "Over-the-air deep learning based radio signal classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, pp. 168-179, 2018.
- [12] S. Hu, Y. Pei, P. P. Liang and Y.-C. Liang, "Robust Modulation Classification Under Uncertain Noise Conditions Using Recurrent Neural Networks," *IEEE Global Communications Conference (GLOBECOM)*, 2018.
- [13] D. Zhang, W. Ding, B. Zhang, C. Xie, C. Liu, J. Han and H. Li, "Heterogeneous Deep Model Fusion for Automatic Modulation Classification," *Preprints*, 1 2018.
- [14] Y. Wang and J. Orchard, "Fast discrete orthonormal Stockwell transform," *SIAM Journal on Scientific Computing*, vol. 31, pp. 4000-4012, 2009.
- [15] R. G. Stockwell, L. Mansinha and R. P. Lowe, "Localization of the complex spectrum: the S transform," *IEEE transactions on signal processing*, vol. 44, pp. 998-1001, 1996.
- [16] I. Daubechies and others, "Ten lectures on wavelets," in *CBMS-NSF regional conference series in applied mathematics*, 1991.
- [17] M. Eramian, R. A. Schincariol, R. G. Stockwell, R. P. Lowe and L. Mansinha, "Review of applications of 1D and 2D S transforms," in *Wavelet Applications IV*, 1997.
- [18] L. Mansinha, R. G. Stockwell, R. P. Lowe, M. Eramian and R. A. Schincariol, "Local S-spectrum analysis of 1-D and 2-D data," *Physics of the Earth and Planetary Interiors*, vol. 103, pp. 329-336, 1997.
- [19] L. Mansinha, R. G. Stockwell and R. P. Lowe, "Pattern analysis with two-dimensional spectral localisation: Applications of two-dimensional S transforms," *Physica A: Statistical Mechanics and its Applications*, vol. 239, pp. 286-295, 1997.
- [20] R. G. Stockwell, "A basis for efficient representation of the S-transform," *Digital Signal Processing*, vol. 17, pp. 371-393, 2007.
- [21] DeepSig, "RF Datasets for Machine Learning," DeepSig Inc., [Online]. Available: <https://www.deepsig.io/datasets/>.
- [22] F. Chollet, "Keras: The Python Deep Learning library," Keras, 2015. [Online]. Available: <https://keras.io/>.

Sparse Bayesian Learning (SBL)-Based Frequency-Selective Channel Estimation for Millimeter Wave Hybrid MIMO Systems

Suraj Srivastava, *Student Member, IEEE*, Ch Suraj Kumar Patro, *Student Member, IEEE*,
Aditya K. Jagannatham, *Member, IEEE*, Govind Sharma, *Member, IEEE*

Abstract—This work develops a novel sparse Bayesian learning (SBL)-based channel estimation technique for frequency-selective millimeter wave (mmWave) multiple-input multiple-output (MIMO) systems. Towards this end, the concatenated frequency-selective MIMO channel matrix is represented in terms of the beamspace channel vector employing suitable transmit and receive array response dictionary matrices. Subsequently, a multiple measurement vector (MMV) model is developed for estimation of the sparse beamspace channel vector considering the block transmission of zero-padded training frames. The unique aspects of the proposed scheme are that it exploits the *group-sparsity* inherent in the equivalent beamspace channel vector of the frequency-selective mmWave MIMO channel and also considers the effect of correlated noise in the equivalent system model due to RF-combining. This feature, coupled with the improved ability of SBL for sparse signal recovery, leads to a significantly enhanced performance of the proposed scheme in comparison to the orthogonal matching pursuit (OMP) technique proposed recently. Bayesian Cramér-Rao bounds (BCRBs) are also derived to characterize the estimation performance. Simulation results are presented to demonstrate the improved performance of the proposed SBL-based channel estimation technique in comparison to the existing scheme and also a performance close to the various benchmarks.

I. INTRODUCTION

Millimeter wave (mmWave) wireless technology, which leverages the vast spectrum available in the 30GHz-300GHz mmWave band, has shown significant potential towards realizing higher data rates in 5G wireless networks [1]. However, communication in the mmWave band is challenging due to the higher path losses, increased hardware complexity and severe signal blockage at such high frequencies [1]. In this context, hybrid beamforming techniques that require a significantly lower number of radio frequency (RF) chains in comparison to the number of transmit/ receive antennas have been shown to be ideally suited for mmWave MIMO transceivers [1], [2]. In such systems, RF precoding/ combining is performed in the RF front-end of the transceiver employing only phase shifters, which yields a beamforming gain. The performance gain is further enhanced via digital precoding/ combining in the baseband, which results in a multiplexing gain. However, the performance of the baseband and RF precoders and combiners designed is critically dependent on the accuracy of available mmWave MIMO channel estimates [1], [2]. Hence,

S. Srivastava, S. K. Patro, A. K. Jagannatham and G. Sharma are with the Department of Electrical Engineering, Indian Institute of Technology Kanpur, Kanpur, India-208016, India (e-mail: ssrivast@iitk.ac.in, chsuraj@iitk.ac.in, adityaj@iitk.ac.in, govind@iitk.ac.in). This work was supported in part by the Qualcomm Innovation Fellowship India-2018 program.

channel estimation plays a central role in realizing the high performance gains promised by mmWave MIMO systems. Therefore, this has naturally been the focus of several research works in this area such as [1]–[5]. A brief review of the existing works in this context is presented next.

A. Review of Existing Works

Several works in the existing literature, such as [2], [3], [6], have considered a frequency flat mmWave MIMO channel model. However, the mmWave MIMO channel, due to its wide bandwidth, is frequency-selective in nature. A few recent works have focused their attention towards developing channel estimation schemes for frequency-selective mmWave MIMO channels. In this context, the time-domain channel estimation approach for single carrier mmWave systems described in [4] employs a spatial grid-based orthogonal matching pursuit (OMP) scheme. However, the scheme proposed therein is sensitive to the choice of the dictionary matrix as well as the stopping criterion, with minor variations leading to convergence errors and performance degradation. Further, it does not consider the effect of the correlated noise arising in the equivalent system model. Moreover, the scheme proposed in [4] does not leverage the *group-sparsity* inherent in the frequency-selective mmWave MIMO channel. Another frequency-domain approach for a mmWave MIMO-OFDM system has been proposed in [5]. The scheme therein aims at estimating the MIMO channel at every sub-carrier and is again based on the OMP scheme. However, the design of hybrid precoders/ combiners for OFDM transmission is challenging because all the sub-carriers have to share a common RF precoder and combiner [5], [7], [8], which might be sub-optimal. Further, since linear power amplifiers are expensive and difficult to design for such large bandwidth systems, it is also desirable for mmWave transmitters to have a low peak-to-average power ratio (PAPR), which favours single carrier (SC) transmission in the mmWave band [9]. Thus there is a need to develop novel channel estimation schemes for single-carrier, frequency-selective mmWave MIMO channels, which overcome the shortcomings of the techniques in the existing literature. A brief summary of the contributions of the paper is presented next.

B. Contributions of the Work

Motivated by the success and improved performance of sparse Bayesian Learning (SBL) technique [10] for sparse signal recovery, in comparison to other existing schemes such

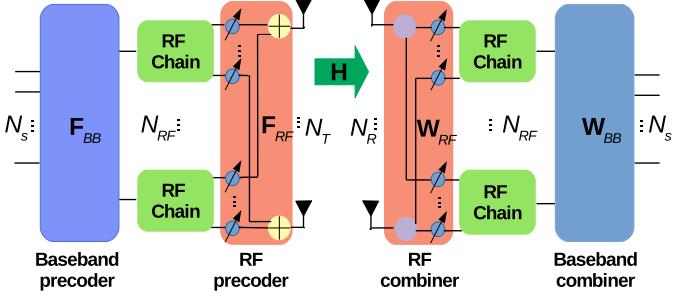


Fig. 1. Schematic Diagram of Hybrid Signal Processing in mmWave MIMO Systems.

as FOCUS (FOcal Underdetermined System Solver), BP (Basis Pursuit) and MP (Matching Pursuit) based algorithms [11], [12], this work proposes a novel SBL-based channel estimation approach for frequency-selective mmWave MIMO channels. Further, since none of the existing works in the literature leverage the *group-sparsity* inherent in the frequency-selective mmWave MIMO channel, this work aims to fill the void by suitably recasting the channel estimation paradigm for the same. The concatenated frequency-selective mmWave MIMO channel matrix has been represented in terms of the *group-sparse* beamspace channel vector employing a suitable sparsifying dictionary that is composed of quantized transmit and receive array response matrices. A multiple measurement vector (MMV) model is then developed for estimation of the sparse beamspace channel vector. The Bayesian Cramér-Rao lower bound (BCRB) is also developed to characterize the efficiency of the proposed *group-SBL* (G-SBL) based estimation scheme. Simulation results are presented to illustrate and compare the performance of the proposed scheme with that of the existing schemes and other benchmarks.

Notation: Boldface lower case and upper case letters denote column vectors and matrices respectively, while \mathbf{I}_N denotes the $N \times N$ identity matrix. For a matrix \mathbf{A} , $\mathbf{A}(i, j)$ denotes the $(i, j)^{th}$ element of \mathbf{A} and similarly $\mathbf{a}(i)$ denotes the i^{th} element of the column vector \mathbf{a} . $\text{vec}(\mathbf{A})$ denotes the vector equivalent of the matrix \mathbf{A} formed by stacking its columns as a single column vector and $\text{vec}^{-1}(\mathbf{a})$ denotes a matrix obtained by the corresponding inverse vectorization operation. The quantities $\text{diag}(a_1, a_2, \dots, a_N)$ and $\text{blkdiag}(\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N)$ denote diagonal and block-diagonal matrices, respectively, with principal diagonal elements being a_1, a_2, \dots, a_N and $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N$. Superscripts $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^*$ and $(\cdot)^{-1}$ denote the transpose, Hermitian, conjugation and inverse operators, respectively. Statistical expectation is represented by $\mathbb{E}\{\cdot\}$, the matrix trace operator by $\text{Tr}(\cdot)$, the matrix determinant operator by $\det(\cdot)$ and the matrix Kronecker product by \otimes . The l_2 norm and Frobenius norm are denoted by $\|\cdot\|_2$ and $\|\cdot\|_F$, respectively.

II. MMWAVE MIMO SYSTEM AND CHANNEL MODEL

A. mmWave MIMO System Model

Consider a mmWave hybrid MIMO system with N_T transmit antennas, N_R receive antennas and $N_{RF} \leq \min(N_T, N_R)$ RF chains, each at the transmitter and receiver, to support

$N_s \leq N_{RF}$ parallel data streams. The hybrid MIMO architecture is shown in Fig. 1. The hybrid precoder $\mathbf{F} = \mathbf{F}_{RF}\mathbf{F}_{BB} \in \mathbb{C}^{N_T \times N_s}$, comprises of a cascade of digital MIMO baseband and analog RF domain precoders, denoted by matrices, $\mathbf{F}_{BB} \in \mathbb{C}^{N_{RF} \times N_s}$ and $\mathbf{F}_{RF} \in \mathbb{C}^{N_T \times N_{RF}}$, respectively. Consider a frequency-selective mmWave MIMO channel with L delay taps between the transmitter and receiver. This can be represented as $\mathbf{H}_d \in \mathbb{C}^{N_R \times N_T}$, $d = 0, 1, \dots, L-1$, where d indicates the tap index. The received signal vector $\mathbf{r}[n] \in \mathbb{C}^{N_R \times 1}$ at the n^{th} time instant in the system described above is given as

$$\mathbf{r}[n] = \sqrt{\rho} \sum_{d=0}^{L-1} \mathbf{H}_d \mathbf{F} \mathbf{s}[n-d] + \mathbf{v}[n], \quad (1)$$

where ρ represents the received signal power, $\mathbf{s}[n] \in \mathbb{C}^{N_s \times 1}$ denotes the transmit symbol vector and $\mathbf{v}[n] \in \mathbb{C}^{N_R \times 1} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_R})$ denotes the circularly symmetric complex additive white Gaussian noise (AWGN) vector with zero mean and covariance matrix $\sigma^2 \mathbf{I}_{N_R}$. The receiver employs the hybrid combiner $\mathbf{W} = \mathbf{W}_{RF}\mathbf{W}_{BB} \in \mathbb{C}^{N_R \times N_s}$ to obtain an N_s -dimensional processed signal as

$$\mathbf{y}[n] = \sqrt{\rho} \sum_{d=0}^{L-1} \mathbf{W}^H \mathbf{H}_d \mathbf{F} \mathbf{s}[n-d] + \mathbf{W}^H \mathbf{v}[n], \quad (2)$$

where the matrices $\mathbf{W}_{BB} \in \mathbb{C}^{N_{RF} \times N_s}$ and $\mathbf{W}_{RF} \in \mathbb{C}^{N_R \times N_{RF}}$ denote the baseband and RF combiners, respectively. It is important to note that, since \mathbf{F}_{RF} and \mathbf{W}_{RF} are implemented using analog phase-shifters, their elements are constrained to have equal norm.

B. Sparse mmWave MIMO Channel Model

Using the parametric model for a mmWave MIMO channel described in [1]–[5], the d^{th} delay tap \mathbf{H}_d of the frequency-selective mmWave MIMO channel can be expressed as

$$\mathbf{H}_d = \sqrt{\frac{N_T N_R}{N_p}} \sum_{\ell=1}^{N_p} \alpha_\ell p(dT_s - \tau_\ell) \mathbf{a}_R(\phi_\ell) \mathbf{a}_T^H(\theta_\ell), \quad (3)$$

where $\alpha_\ell \in \mathbb{C}$ and $\tau_\ell \in \mathbb{R}$ represent the complex channel gain and delay, respectively, of the ℓ^{th} spatial multipath component, and $p(\tau)$ denotes the band-limited pulse shaping filter response evaluated at τ . The quantities $\phi_\ell \in [0, \pi)$ and $\theta_\ell \in [0, \pi)$ are the angles of arrival and departure, respectively, of the ℓ^{th} multipath component and $\mathbf{a}_R(\phi_\ell) \in \mathbb{C}^{N_R \times 1}$ and $\mathbf{a}_T(\theta_\ell) \in \mathbb{C}^{N_T \times 1}$ denote the array response vectors corresponding to uniform linear receive and transmit arrays, respectively, which are given as

$$\mathbf{a}_R(\phi_\ell) = \frac{1}{\sqrt{N_R}} \left[1, e^{-j \frac{2\pi}{\lambda} d_R \cos \phi_\ell}, \dots, e^{-j \frac{2\pi}{\lambda} (N_R-1) d_R \cos \phi_\ell} \right]^T, \quad (4)$$

$$\mathbf{a}_T(\theta_\ell) = \frac{1}{\sqrt{N_T}} \left[1, e^{-j \frac{2\pi}{\lambda} d_T \cos \theta_\ell}, \dots, e^{-j \frac{2\pi}{\lambda} (N_T-1) d_T \cos \theta_\ell} \right]^T. \quad (5)$$

The quantities λ , d_R and d_T above denote the operating wavelength, receive and transmit antenna spacings, respectively. The channel model in (3) can be succinctly represented as

$$\mathbf{H}_d = \mathbf{A}_R \mathbf{D}_d \mathbf{A}_T^H, \quad (6)$$

where $\mathbf{D}_d \in \mathbb{C}^{N_p \times N_p}$ is a diagonal matrix with entries $\sqrt{\frac{N_T N_R}{N_p}} \alpha_\ell p(dT_s - \tau_\ell)$, $\ell = 1, 2, \dots, N_p$, on its principal diagonal and $\underline{\mathbf{A}}_R \in \mathbb{C}^{N_R \times N_p}$ and $\underline{\mathbf{A}}_T \in \mathbb{C}^{N_T \times N_p}$ denote the concatenated matrices of the receive and transmit array response vectors, $\mathbf{a}_R(\phi_\ell)$ and $\mathbf{a}_T(\theta_\ell)$, $\ell = 1, 2, \dots, N_p$, respectively. The sparse channel model for the mmWave hybrid MIMO system is developed next.

Consider a partition of the feasible angle of departure (AoD) and angle of arrival (AoA) space $[0, \pi]$ with $G_T, G_R \geq \max\{N_T, N_R\}$ angular grid points. The sets of quantized spatial angles $\Phi_R = \{\phi_g : \phi_g \in [0, \pi], \forall 1 \leq g \leq G_R\}$ and $\Theta_T = \{\theta_g : \theta_g \in [0, \pi], \forall 1 \leq g \leq G_T\}$, corresponding to the transmit and receive antenna arrays, respectively, are chosen such that the following condition is satisfied [3]

$$\cos(\phi_g) = \frac{2}{G_R}(g-1) - 1, \forall 1 \leq g \leq G_R, \quad (7)$$

$$\cos(\theta_g) = \frac{2}{G_T}(g-1) - 1, \forall 1 \leq g \leq G_T. \quad (8)$$

The transmit and receive array response dictionary matrices $\mathbf{A}_T(\Theta_T) = [\mathbf{a}_T(\theta_1), \dots, \mathbf{a}_T(\theta_{G_T})]$ and $\mathbf{A}_R(\Phi_R) = [\mathbf{a}_R(\phi_1), \dots, \mathbf{a}_R(\phi_{G_R})]$ are obtained by concatenating the array response vectors corresponding to the angular grids Θ_T and Φ_R , respectively. From (7) and (8), the transmit and receive array response matrices can also be seen to satisfy the conditions $\mathbf{A}_T(\Theta_T)\mathbf{A}_T^H(\Theta_T) = \frac{G_T}{N_T}\mathbf{I}_{N_T}$ and $\mathbf{A}_R(\Phi_R)\mathbf{A}_R^H(\Phi_R) = \frac{G_R}{N_R}\mathbf{I}_{N_R}$ [3]. The beamspace representation of the channel matrix $\mathbf{H}_d \in \mathbb{C}^{N_R \times N_T}$ can now be obtained in terms of the array response dictionary matrices $\mathbf{A}_T(\Theta_T)$ and $\mathbf{A}_R(\Phi_R)$ as

$$\mathbf{H}_d = \mathbf{A}_R(\Phi_R)\mathbf{H}_{b,d}\mathbf{A}_T^H(\Theta_T), \quad (9)$$

where $\mathbf{H}_{b,d} \in \mathbb{C}^{G_R \times G_T}$ denotes the equivalent beamspace channel matrix corresponding to channel tap matrix \mathbf{H}_d . The vector representation of the channel matrix above is obtained by stacking the columns of \mathbf{H}_d as

$$\mathbf{h}_d = \text{vec}(\mathbf{H}_d) = (\mathbf{A}_T^H(\Theta_T) \otimes \mathbf{A}_R(\Phi_R)) \mathbf{h}_{b,d}, \quad (10)$$

where $\mathbf{h}_{b,d} \triangleq \text{vec}(\mathbf{H}_{b,d}) \in \mathbb{C}^{G_R G_T \times 1}$ represents the equivalent beamspace channel vector obtained by a column wise stacking of the beamspace matrix $\mathbf{H}_{b,d}$. As described in related works such as [2], [3] on mmWave MIMO systems, only a few components $N_p (< G_R G_T)$ of the beamspace channel vector $\mathbf{h}_{b,d}$ are active (non zero), owing to the highly directional nature of signal propagation at mmWave frequencies coupled with the reduced multipath scattering effect. The quantity N_p , which denotes the number of scatterers in the mmWave MIMO channel, corresponds to the number of active transmit and receive directional cosine pairs. Thus, the unknown beamspace channel vector $\mathbf{h}_{b,d}$ in (10) is N_p -sparse.

For the purpose of channel estimation, the concatenated vector equivalent channel $\mathbf{h} \in \mathbb{C}^{N_R N_T L \times 1}$ for the mmWave

MIMO frequency selective channel can be constructed as

$$\begin{aligned} \mathbf{h} &\triangleq \text{vec}(\underbrace{[\mathbf{H}_0 \ \mathbf{H}_1 \ \dots \ \mathbf{H}_{L-1}]}_{\mathbf{H}}) = [\mathbf{h}_0^T \ \mathbf{h}_1^T \ \dots \ \mathbf{h}_{L-1}^T]^T, \\ &= \underbrace{(\mathbf{I}_L \otimes \mathbf{A}_T^*(\Theta_T) \otimes \mathbf{A}_R(\Phi_R))}_{\Psi} \mathbf{h}_b, \end{aligned} \quad (11)$$

where $\mathbf{h}_b \triangleq [\mathbf{h}_{b,0}^T \ \mathbf{h}_{b,1}^T \ \dots \ \mathbf{h}_{b,L-1}^T]^T \in \mathbb{C}^{G_R G_T L \times 1}$ is the vector equivalent beamspace channel that is *group-sparse*¹ due to natural grouping of its components. The next section develops the framework for sparse channel estimation.

III. SPARSE CHANNEL ESTIMATION MODEL

For the purpose of channel estimation, consider the transmission of a block of M training frames, each of length N . A zero-prefix (ZP) of length $L-1$ is added to each frame. The transmitter and receiver employ the RF precoder $\mathbf{F}_{RF,m} \in \mathbb{C}^{N_T \times N_{RF}}$ and RF combiner $\mathbf{W}_{RF,m} \in \mathbb{C}^{N_R \times N_{RF}}$, respectively, during the m^{th} training frame. Let $\mathbf{s}_m[n] \in \mathbb{C}^{N_{RF} \times 1}$ denote the pilot vector at time instant n in the m^{th} training frame. The corresponding RF combiner output is given as

$$\mathbf{y}_m[n] = \sqrt{\rho} \sum_{d=0}^{L-1} \mathbf{W}_{RF,m}^H \mathbf{H}_d \mathbf{F}_{RF,m} \mathbf{s}_m[n-d] + \mathbf{W}_{RF,m}^H \mathbf{v}_m[n], \quad (12)$$

where $\mathbf{v}_m[n]$, similar to $\mathbf{v}[n]$ in equation (1) and (2), is $\mathcal{CN}(0, \sigma^2 \mathbf{I}_{N_R})$. Let $\mathbf{Y}_m \in \mathbb{C}^{N_{RF} \times N}$ be obtained by concatenating the N -measurements in frame m as $\mathbf{Y}_m = [\mathbf{y}_m[1] \ \mathbf{y}_m[2] \ \dots \ \mathbf{y}_m[N]]$. One can express the resulting multiple measurement vector (MMV) model as

$$\mathbf{Y}_m = \sqrt{\rho} \mathbf{W}_{RF,m}^H \mathbf{H} (\mathbf{I}_L \otimes \mathbf{F}_{RF,m}) \mathbf{S}_m^T + \mathbf{E}_m, \quad (13)$$

where $\mathbf{E}_m \triangleq [\mathbf{W}_{RF,m}^H \mathbf{v}_m[1] \ \dots \ \mathbf{W}_{RF,m}^H \mathbf{v}_m[N]] \in \mathbb{C}^{N_{RF} \times N}$ is the concatenated RF combiner output noise matrix and the training symbol matrix $\mathbf{S}_m \in \mathbb{C}^{N \times N_{RF} L}$ is defined as

$$\mathbf{S}_m \triangleq \begin{bmatrix} \mathbf{s}_m^T[1] & 0 & \dots & 0 \\ \mathbf{s}_m^T[2] & \mathbf{s}_m^T[1] & \dots & \cdot \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{s}_m^T[N] & \dots & \dots & \mathbf{s}_m^T[N-L+1] \end{bmatrix}.$$

An equivalent model is derived for the system above by stacking the columns of \mathbf{Y}_m as

$$\begin{aligned} \mathbf{y}_m &\triangleq \text{vec}(\mathbf{Y}_m) \in \mathbb{C}^{N N_{RF} \times 1} \\ &= \underbrace{\sqrt{\rho} \mathbf{S}_m (\mathbf{I}_L \otimes \mathbf{F}_{RF,m}^T) \otimes \mathbf{W}_{RF,m}^H}_{\Phi_m} \mathbf{h} + \mathbf{e}_m, \end{aligned} \quad (14)$$

where $\mathbf{e}_m \triangleq \text{vec}(\mathbf{E}_m) \in \mathbb{C}^{N N_{RF} \times 1}$ is the stacked noise vector with covariance matrix $\mathbb{E}[\mathbf{e}_m \mathbf{e}_m^H] = \sigma^2 (\mathbf{I}_N \otimes \mathbf{W}_{RF,m}^H \mathbf{W}_{RF,m}) \in \mathbb{C}^{N N_{RF} \times N N_{RF}}$. Substituting (11) in (14), the equivalent system model for estimation of the beamspace channel vector \mathbf{h}_b from the m^{th} training frame can be represented in a compact fashion as

$$\mathbf{y}_m = \Phi_m \Psi \mathbf{h}_b + \mathbf{e}_m. \quad (15)$$

¹The components within each group of a *group-sparse* vector are likely to be either all zeros or all nonzeros.

The aggregate system model over the M training frames can be expressed as

$$\mathbf{y} = \Phi \mathbf{h}_b + \mathbf{e}, \quad (16)$$

where $\mathbf{y} \triangleq [\mathbf{y}_1^T \ \mathbf{y}_2^T \cdots \mathbf{y}_M^T]^T \in \mathbb{C}^{NMN_{\text{RF}} \times 1}$ and $\mathbf{e} \triangleq [\mathbf{e}_1^T \ \mathbf{e}_2^T \cdots \mathbf{e}_M^T]^T \in \mathbb{C}^{NMN_{\text{RF}} \times 1}$ are the stacked measurement and noise vectors for the M -training frames. The matrix $\Phi \in \mathbb{C}^{NMN_{\text{RF}} \times G_R G_T L}$ denotes the equivalent dictionary matrix that is defined as $\Phi \triangleq [\Phi_1^T \ \Phi_2^T \cdots \Phi_M^T]^T \Psi$. The procedure to construct the RF precoder and combiner matrices, $\mathbf{F}_{\text{RF},m}$ and $\mathbf{W}_{\text{RF},m}$, respectively, for the M training frames is described in section-V. Finally, the block diagonal covariance matrix $\mathbf{R}_e \triangleq \mathbb{E}[\mathbf{e}\mathbf{e}^H] \in \mathbb{C}^{NMN_{\text{RF}} \times NMN_{\text{RF}}}$ of the noise vector \mathbf{e} in (16) is given as

$$\mathbf{R}_e = \sigma^2 \text{blkdiag} \left(\{\mathbf{I}_N \otimes \mathbf{W}_{\text{RF},m}^H \mathbf{W}_{\text{RF},m}\}_{m=1}^M \right). \quad (17)$$

The next section presents the *group-SBL* (G-SBL) framework for estimation of the beamspace channel vector \mathbf{h}_b that leverages its inherent *group-sparsity* for improved estimation performance.

IV. G-SBL FOR GROUP-SPARSE MMWAVE MIMO CHANNEL ESTIMATION

The G-SBL framework begins by assigning the parameterized Gaussian prior below to the unknown beamspace channel vector corresponding to the d^{th} delay tap $\mathbf{h}_{b,d}$ [10]

$$p(\mathbf{h}_{b,d}; \Gamma) = \prod_{i=1}^{G_R G_T} (\pi \gamma_i)^{-1} \exp \left(-\frac{|\mathbf{h}_{b,d}(i)|^2}{\gamma_i} \right), \quad (18)$$

where the hyperparameter $\gamma_i, \forall 1 \leq i \leq G_R G_T$ denotes the prior variance corresponding to the i^{th} component of $\mathbf{h}_{b,d}$ and $\Gamma \triangleq \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_{G_R G_T}) \in \mathbb{R}^{G_R G_T \times G_R G_T}$ is the diagonal matrix of hyperparameters. The G-SBL based parameterized prior assignment for the concatenated beamspace vector \mathbf{h}_b is determined as

$$p(\mathbf{h}_b; \Gamma) = \prod_{d=0}^{L-1} \prod_{i=1}^{G_R G_T} (\pi \gamma_i)^{-1} \exp \left(-\frac{|\mathbf{h}_{b,d}(i)|^2}{\gamma_i} \right). \quad (19)$$

The *a posteriori* probability density function of the beamspace channel vector \mathbf{h}_b can be evaluated as $p(\mathbf{h}_b | \mathbf{y}; \Gamma) \sim \mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where the quantities $\boldsymbol{\mu} \in \mathbb{C}^{G_R G_T L \times 1}$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{G_R G_T L \times G_R G_T L}$ are obtained as

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \boldsymbol{\Phi}^H \mathbf{R}_e^{-1} \mathbf{y}, \quad \boldsymbol{\Sigma} = \left(\boldsymbol{\Phi}^H \mathbf{R}_e^{-1} \boldsymbol{\Phi} + (\mathbf{I}_L \otimes \Gamma)^{-1} \right)^{-1}, \quad (20)$$

which can be seen to depend on the hyperparameter matrix Γ . Thus, estimation of the beamspace channel vector \mathbf{h}_b reduces to the estimation of the associated hyperparameter vector $\boldsymbol{\gamma} \triangleq [\gamma_1, \gamma_2, \dots, \gamma_{G_R G_T}]^T$. The G-SBL approach chooses the hyperparameter matrix $\widehat{\Gamma}$ and in turn the prior $p(\mathbf{h}_b; \widehat{\Gamma})$ that maximizes the Bayesian evidence $p(\mathbf{y}; \Gamma)$, thus leading to an improved performance for estimation of the sparse vector \mathbf{h}_b .

The Bayesian evidence $p(\mathbf{y}; \Gamma)$ can now be maximized using the iterative expectation-maximization (EM) algorithm. Let $\widehat{\Gamma}^{(k-1)}$ denote the estimate of the hyperparameter matrix Γ in the $(k-1)^{\text{th}}$ iteration. The expectation (E-step) in the

k^{th} iteration evaluates the average log-likelihood $\mathcal{L}(\Gamma | \widehat{\Gamma}^{(k-1)})$ of the complete data as

$$\begin{aligned} \mathcal{L}(\Gamma | \widehat{\Gamma}^{(k-1)}) &= \mathbb{E}_{\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}} \left\{ \log p(\mathbf{y}, \mathbf{h}_b; \Gamma) \right\} \\ &= \mathbb{E}_{\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}} \left\{ \log p(\mathbf{y} | \mathbf{h}_b) + \log p(\mathbf{h}_b; \Gamma) \right\}. \end{aligned} \quad (21)$$

Subsequently, the maximization (M-step), which maximizes $\mathcal{L}(\Gamma | \widehat{\Gamma}^{(k-1)})$ with respect to the hyperparameter vector $\boldsymbol{\gamma}$ can be represented as

$$\widehat{\boldsymbol{\gamma}}^{(k)} = \arg \max_{\boldsymbol{\gamma}} \mathbb{E}_{\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}} \left\{ \log p(\mathbf{y} | \mathbf{h}_b) + \log p(\mathbf{h}_b; \Gamma) \right\}. \quad (22)$$

The first term inside the $\mathbb{E}\{\cdot\}$ operator above, which simplifies as $-MN N_{\text{RF}} \log \pi - \log \det(\mathbf{R}_e) - (\mathbf{y} - \Phi \mathbf{h}_b)^H \mathbf{R}_e^{-1} (\mathbf{y} - \Phi \mathbf{h}_b)$, is seen to be independent of the hyperparameter vector $\boldsymbol{\gamma}$, and can thus be ignored in the maximization. The equivalent optimization problem for estimation of the hyperparameter vector $\boldsymbol{\gamma}$ is given as

$$\begin{aligned} \widehat{\boldsymbol{\gamma}}^{(k)} &\equiv \arg \max_{\boldsymbol{\gamma}} \mathbb{E}_{\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}} \left\{ \log p(\mathbf{h}_b; \Gamma) \right\} \\ &\equiv \arg \max_{\boldsymbol{\gamma}} \sum_{i=1}^{G_R G_T} \left(-L \log \gamma_i + \sum_{d=0}^{L-1} -\frac{\mathbb{E}_{\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}} \left\{ |\mathbf{h}_{b,d}(i)|^2 \right\}}{\gamma_i} \right). \end{aligned} \quad (23)$$

It can be readily observed that the maximization problem above towards estimation of the hyperparameter vector $\boldsymbol{\gamma}$ is decoupled with respect to each γ_i . This can therefore be solved to obtain the estimates $\widehat{\gamma}_i^{(k)}$ in the k^{th} iteration of the EM algorithm as

$$\begin{aligned} \widehat{\gamma}_i^{(k)} &= \frac{1}{L} \sum_{d=0}^{L-1} \mathbb{E}_{\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}} \left\{ |\mathbf{h}_{b,d}(i)|^2 \right\} \\ &= \frac{1}{L} \sum_{d=0}^{L-1} \boldsymbol{\Sigma}^{(k)}(\tilde{d}, \tilde{d}) + |\boldsymbol{\mu}^{(k)}(\tilde{d})|^2, \end{aligned} \quad (24)$$

where $\tilde{d} = d G_R G_T + i$. The estimate of the hyperparameter matrix follows as $\boldsymbol{\Gamma}^{(k)} = \text{diag}(\boldsymbol{\gamma}^{(k)})$. The estimates of the mean $\boldsymbol{\mu}^{(k)}$ and covariance matrix $\boldsymbol{\Sigma}^{(k)}$ corresponding to the *a posteriori* probability density function $p(\mathbf{h}_b | \mathbf{y}; \widehat{\Gamma}^{(k-1)}) \sim \mathcal{CN}(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ can be obtained by substituting $\boldsymbol{\Gamma} = \widehat{\Gamma}^{(k-1)}$ in (20). On convergence, the G-SBL estimate of the beamspace channel vector is determined as the posterior mean, i.e., $\widehat{\mathbf{h}}_b = \boldsymbol{\mu}^{(k)}$. A concise summary of the various steps in the proposed G-SBL technique for sparse channel estimation is given in Algorithm 1. The G-SBL based estimate of the matrix $\widehat{\mathbf{H}}_d$, which corresponds to the d^{th} tap of the frequency selective mmWave MIMO channel, is obtained as

$$\widehat{\mathbf{H}}_d = \mathbf{A}_R(\boldsymbol{\Phi}_R) \text{vec}^{-1} \left(\widehat{\mathbf{h}}_{b,d} \right) \mathbf{A}_T^H(\boldsymbol{\Theta}_T), \quad (25)$$

where $\widehat{\mathbf{h}}_{b,d} \triangleq \widehat{\mathbf{h}}_b(d G_R G_T + 1 : (d+1) G_R G_T)$. The next subsection presents the BCRB for the mean square error (MSE) of the frequency selective *group-sparse* mmWave MIMO channel estimate.

Algorithm 1 G-SBL based frequency selective mmWave channel estimation

Input: Observation $\mathbf{y} \in \mathbb{C}^{NMN_{\text{RF}} \times 1}$, Dictionary Matrix $\Phi \in \mathbb{C}^{NMN_{\text{RF}} \times G_R G_T L}$, Noise Covariance $\mathbf{R}_e \in \mathbb{C}^{NMN_{\text{RF}} \times NMN_{\text{RF}}}$, Stopping Parameters ϵ and k_{\max}

Initialization: $\hat{\gamma}_i^{(0)} = 1, \forall 1 \leq i \leq G_R G_T \implies \hat{\Gamma}^{(0)} = \mathbf{I}$

Set counter $k = 0$ and $\hat{\Gamma}^{(-1)} = \mathbf{0}$

while $\|\hat{\gamma}^{(k)} - \hat{\gamma}^{(k-1)}\|_2 > \epsilon$ and $k < k_{\max}$ **do**

$$k \leftarrow k + 1$$

E-step: Evaluate *a posteriori* probability density $p(\mathbf{h}_b | \mathbf{y}; \hat{\Gamma}^{(k-1)}) \sim \mathcal{CN}(\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$

$$\boldsymbol{\Sigma}^{(k)} = (\Phi^H \mathbf{R}_e^{-1} \Phi + (\mathbf{I}_L \otimes \hat{\Gamma}^{(k-1)})^{-1})^{-1}$$

$$\boldsymbol{\mu}^{(k)} = \boldsymbol{\Sigma}^{(k)} \Phi^H \mathbf{R}_e^{-1} \mathbf{y}$$

M-step: Evaluate hyperparameter estimates

for $i = 0, 1, \dots, G_R G_T$ **do**

$$\tilde{d} = dG_R G_T + i$$

$$\hat{\gamma}_i^{(k)} = \frac{1}{L} \sum_{d=0}^{L-1} \boldsymbol{\Sigma}^{(k)}(\tilde{d}, \tilde{d}) + |\boldsymbol{\mu}^{(k)}(\tilde{d})|^2$$

end for

end while

Output: $\hat{\mathbf{h}}_b = \boldsymbol{\mu}^{(k)}$

A. Bayesian Cramér-Rao Lower Bound

The Bayesian Fisher information matrix (BFIM) \mathbf{J}_B for the beamspace channel vector \mathbf{h}_b can be expressed as

$$\mathbf{J}_B = \mathbf{J}_D + \mathbf{J}_P, \quad (26)$$

where $\mathbf{J}_D \in \mathbb{C}^{G_R G_T L \times G_R G_T L}$ and $\mathbf{J}_P \in \mathbb{C}^{G_R G_T L \times G_R G_T L}$ denote the FIMs corresponding to the measurement vector \mathbf{y} and beamspace channel vector \mathbf{h}_b , respectively, which are determined as

$$\mathbf{J}_D = -\mathbb{E}_{(\mathbf{y}, \mathbf{h}_b)} \left\{ \frac{\partial^2 \mathcal{L}(\mathbf{y} | \mathbf{h}_b)}{\partial \mathbf{h}_b \partial \mathbf{h}_b^H} \right\}, \quad \mathbf{J}_P = -\mathbb{E}_{\mathbf{h}_b} \left\{ \frac{\partial^2 \mathcal{L}(\mathbf{h}_b; \boldsymbol{\Gamma})}{\partial \mathbf{h}_b \partial \mathbf{h}_b^H} \right\}. \quad (27)$$

The log-likelihood of the measurement and log-prior density of the beamspace channel vectors are $\mathcal{L}(\mathbf{y} | \mathbf{h}_b) = \log p(\mathbf{y} | \mathbf{h}_b)$ and $\mathcal{L}(\mathbf{h}_b; \boldsymbol{\Gamma}) = \log p(\mathbf{h}_b; \boldsymbol{\Gamma})$, respectively, which are obtained as

$$\mathcal{L}(\mathbf{y} | \mathbf{h}_b) = \kappa_1 - (\mathbf{y} - \Phi \mathbf{h}_b)^H \mathbf{R}_e^{-1} (\mathbf{y} - \Phi \mathbf{h}_b), \quad (28)$$

$$\mathcal{L}(\mathbf{h}_b; \boldsymbol{\Gamma}) = \kappa_2 - \mathbf{h}_b^H (\mathbf{I}_L \otimes \boldsymbol{\Gamma})^{-1} \mathbf{h}_b, \quad (29)$$

where the constants $\kappa_1 = -M N N_{\text{RF}} \log \pi - \log \det(\mathbf{R}_e)$ and $\kappa_2 = -G_R G_T L \log(\pi) - \log(\det(\mathbf{I}_L \otimes \boldsymbol{\Gamma}))$. Substituting the above quantities in (27), the matrices \mathbf{J}_D and \mathbf{J}_P can be determined as $\mathbf{J}_D = \Phi^H \mathbf{R}_e^{-1} \Phi$ and $\mathbf{J}_P = (\mathbf{I}_L \otimes \boldsymbol{\Gamma})^{-1}$. Hence, the Bayesian CRB for the MSE of the estimate \mathbf{h}_b is given by

$$\text{MSE}(\hat{\mathbf{h}}_b) \geq \text{Tr}\{\mathbf{J}_B^{-1}\} = \text{Tr}\left\{\left(\Phi^H \mathbf{R}_e^{-1} \Phi + (\mathbf{I}_L \otimes \boldsymbol{\Gamma})^{-1}\right)^{-1}\right\}. \quad (30)$$

From (11), the BCRB for the MSE of the concatenated frequency selective mmWave MIMO channel vector \mathbf{h} is in turn determined as

$$\text{MSE}(\hat{\mathbf{h}}) \geq \text{Tr}\left\{\boldsymbol{\Psi} \left(\boldsymbol{\Psi}^H \mathbf{R}_e^{-1} \boldsymbol{\Psi} + (\mathbf{I}_L \otimes \boldsymbol{\Gamma})^{-1} \right)^{-1} \boldsymbol{\Psi}^H\right\}. \quad (31)$$

V. SIMULATION RESULTS

A mmWave MIMO system is considered with number of transmit and receive antennas set as $N_T = N_R \in \{8, 32\}$ and number of RF chains $N_{\text{RF}} \in \{2, 6\}$. The inter-antenna spacings of the transmit and receive antenna arrays are fixed as $\frac{d_T}{\lambda} = \frac{d_R}{\lambda} = \frac{1}{2}$ and the set of feasible AoA/ AoD space comprises of $G_T = G_R = G \in \{16, 32\}$ angular grid points. The mmWave MIMO channel is assumed to be spatially sparse with $N_p = 4$ scatterers and $L = 4$ delay taps. The corresponding path gains α_ℓ are modeled as i.i.d. $\mathcal{CN}(0, 1/L)$ random variables and the delays τ_ℓ are chosen uniformly at random from $[0, L-1]$. The raised cosine pulse shaping filter is assumed to have a roll-off factor of 0.6 and the training frame length to assist the channel estimation process is set as $N = 8$. The angle quantization employed in the phase shifter is assumed to have $N_q = 8$ -quantization bits and the entries of \mathbf{F}_m and $\mathbf{W}_m, \forall 1 \leq m \leq M$, are drawn with an equal probability from the set below

$$\mathcal{A} = \left\{ 0, \frac{2\pi}{2^{N_q}}, \dots, \frac{(2^{N_q} - 1)2\pi}{2^{N_q}} \right\}, \quad (32)$$

which implies $\mathbf{F}_m(i, j) = \frac{1}{\sqrt{N_T}} e^{j\varphi_{i,j}}$ and $\mathbf{W}_m(i, j) = \frac{1}{\sqrt{N_R}} e^{j\psi_{i,j}}$, with $\varphi_{i,j}, \psi_{i,j} \in \mathcal{A}$. The SNR is defined as $\frac{\rho}{\sigma^2}$. The stopping criterion for the existing OMP-based approach is set similar to [4], i.e., the algorithm terminates when the residual error is lower than the noise floor $\text{Tr}(\mathbf{R}_e)$. For the G-SBL based approach, the initial value of all the hyperparameters is set as $\hat{\gamma}_i^{(0)} = 1, \forall 1 \leq i \leq G^2$ as described in Algorithm 1 with stopping parameters $\epsilon = 10^{-9}$ and $k_{\max} = 100$.

Fig. 2(a) compares the performance of the proposed G-SBL and existing OMP-based mmWave channel estimation approaches in terms of the normalized MSE (NMSE) defined as $\frac{1}{N_T N_R} \|\hat{\mathbf{H}} - \mathbf{H}\|_F^2$ for the mmWave MIMO setup with $N_T = N_R = 32, N_{\text{RF}} = 6$ and $G = 32$. The number of training frames for channel estimation is set as $M = 6$. From the figure, the proposed G-SBL scheme is seen to yield a significant NMSE reduction in comparison to the existing OMP-based technique in [4]. This arises due to the fact that the performance of the OMP technique is sensitive to the choice of dictionary matrix Φ and the stopping criterion, unlike the SBL that has a robust performance. Further, OMP does not leverage the group-sparsity inherent in the beamspace mmWave MIMO channel vector \mathbf{h}_b and also does not consider the effect of correlated noise \mathbf{e} in (16) at the output of the RF-combiner. The performance is also benchmarked against that of the corresponding BCRB developed in section IV-A and the Oracle-LS estimator with a known channel sparsity profile. In contrast to the OMP scheme, the performance of the G-SBL technique is seen to closely approach that of the oracle

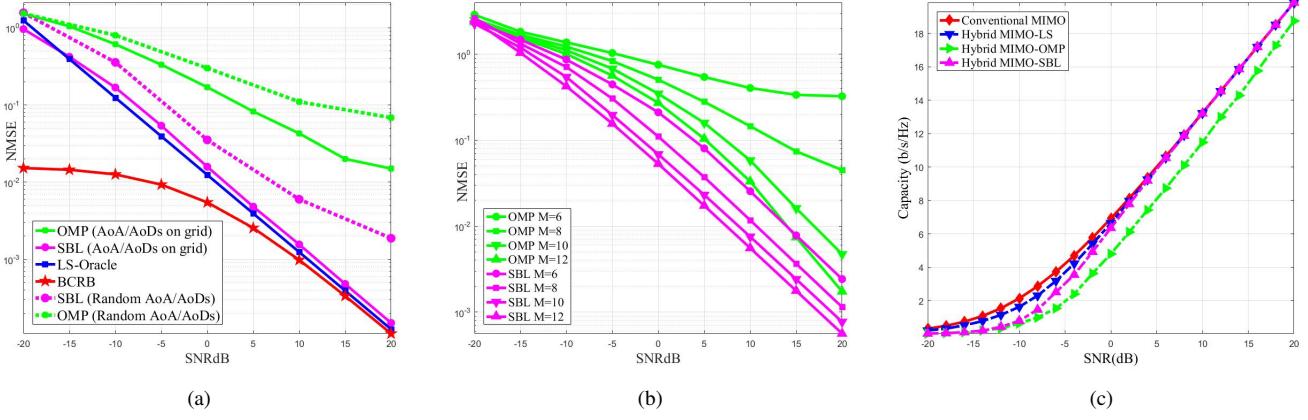


Fig. 2. (a) NMSE vs SNR for the mmWave MIMO setup with $N_T = N_R = 32$, $N_{RF} = 6$, $G = 32$, $N = 8$ and $M = 6$. (b) NMSE vs SNR for the mmWave MIMO setup with $N_T = N_R = 8$, $N_{RF} = 2$, $G = 16$ and $N = 8$. (c) Spectral Efficiency vs SNR for the mmWave MIMO setup with $N_T = N_R = 8$, $N_{RF} = 2$, $G = 16$, $N = 8$ and $M = 6$.

estimator and also the BCRB for higher SNRs, demonstrating its improved capability for sparse channel estimation.

Fig. 2(b) depicts the NMSE performance for the mmWave MIMO system with $N_T = N_R = 8$, $N_{RF} = 2$ and $G = 16$ for a varying number of training frames M from 6 to 12. The NMSE performance of both techniques is seen to improve with increasing the number of training frames thanks to the larger number of measurements. However, it can be observed that the performance of the G-SBL scheme with $M = 6$ training frames is comparable with that of the OMP scheme with $M = 12$. Thus, G-SBL can reduce the training overheads in such systems for a desired level of estimation accuracy.

Fig. 2(c) presents the ergodic rate of the various competing schemes with transmit covariance $\mathbb{E}\{\mathbf{s}[n]\mathbf{s}^H[n]\} = \frac{1}{N_s} \mathbf{I}_{N_s}$. The computation for the ergodic spectral efficiency is provided in [5]. The improved channel estimation accuracy of the G-SBL based scheme is also reflected in its enhanced spectral efficiency. Further, the spectral efficiency of G-SBL scheme is very close to that of the conventional fully digital MIMO precoder/ combiner with perfect CSI.

VI. CONCLUSION

In this work, an SBL-based scheme has been developed for channel estimation in a frequency selective mmWave hybrid MIMO system that leverages the inherent *group-sparsity* using multiple concatenated measurements. The BCRB was also determined to analytically lower bound the MSE performance. Simulation results demonstrated the improved performance in comparison to the existing OMP-based mmWave MIMO channel estimation scheme and also a performance close to that of the genie benchmarks and BCRBs. Further, the improved performance of G-SBL in comparison to OMP can be attributed to its robust ability of sparse signal recovery. Future works can extend the proposed G-SBL and its low complexity alternatives to spatio-temporally correlated frequency selective mmWave MIMO channels.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Raj Kumar and Dr. Shashidhar Vummintala from Qualcomm, India, for their valu-

able comments and suggestions.

REFERENCES

- [1] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 436–453, 2016.
- [2] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, “Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, 2014.
- [3] J. Lee, G.-T. Gil, and Y. H. Lee, “Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications,” *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2370–2386, 2016.
- [4] K. Venugopal, A. Alkhateeb, N. G. Prelicic, and R. W. Heath, “Channel estimation for hybrid architecture-based wideband millimeter wave systems,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1996–2009, 2017.
- [5] J. R. Fernández, N. G. Prelicic, K. Venugopal, and R. W. Heath, “Frequency-domain compressive channel estimation for frequency-selective hybrid millimeter wave MIMO systems,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 2946–2960, 2018.
- [6] A. Mishra, A. Rajoria, A. Jagannatham, and G. Ascheid, “Sparse Bayesian learning-based channel estimation in millimeter wave hybrid MIMO systems,” in *IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2017, pp. 1–6.
- [7] W. Huang, Y. Huang, R. Zhao, S. He, and L. Yang, “Wideband millimeter wave communication: Single carrier based hybrid precoding with sparse optimization,” *IEEE Transactions on Vehicular Technology*, 2018.
- [8] S. Buzzi, C. D Andrea, T. Foggi, A. Ugolini, and G. Colavolpe, “Single-carrier modulation versus OFDM for millimeter-wave wireless MIMO,” *IEEE Transactions on Communications*, vol. 66, no. 3, pp. 1335–1348, 2018.
- [9] J. Mo, P. Schniter, and R. W. Heath, “Channel estimation in broadband millimeter wave mimo systems with few-bit adcs,” *IEEE Transactions on Signal Processing*, vol. 66, no. 5, pp. 1141–1154, 2018.
- [10] D. P. Wipf and B. D. Rao, “Sparse Bayesian learning for basis selection,” *IEEE Transactions on Signal processing*, vol. 52, no. 8, pp. 2153–2164, 2004.
- [11] I. F. Gorodnitsky and B. D. Rao, “Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm,” *IEEE Transactions on signal processing*, vol. 45, no. 3, pp. 600–616, 1997.
- [12] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

MIMO-FBMC Channel Estimation with Limited, and Imperfect Knowledge of Channel Correlations

Prem Singh and K. Vasudevan
Electrical Engineering Department
Indian Institute of Technology Kanpur, India
Email: [psrawat,vasu]@iitk.ac.in

Abstract—This paper presents and analyses the performance of training-based least squares (LS) and minimum mean square error (MMSE) channel estimation schemes for multiple input multiple output (MIMO) filter bank multicarrier (FBMC) systems based on the offset quadrature amplitude modulation (OQAM) in the presence of limited, and imperfect knowledge of the channel correlations. First, a linear MMSE (LMMSE) technique for MIMO-FBMC channel estimation, which require a priori knowledge of channel correlation matrix, is examined by utilizing the second-order statistical properties of the intrinsic interference in FBMC systems. A biased LS (BLS) and relaxed LMMSE (RLMMSE) MIMO-FBMC channel estimation schemes, which require prior knowledge of the trace of the channel correlation matrix, are proposed. The LS-BLS and LS-RLMMSE schemes for MIMO-FBMC channel estimation are investigated in the presence of imperfect knowledge of the channel correlations. The mean square error is derived for the proposed schemes by exploiting statistical properties of the intrinsic interference. Simulation results show that the proposed schemes present an excellent trade-off between the achieved performance and required a priori knowledge of the channel correlations.

I. INTRODUCTION

Orthogonal frequency division multiplexing (OFDM) is a widely popular signaling scheme for broadband wired and wireless communication systems. The ability of OFDM to partition the wideband spectrum into multiple subbands with orthogonal subcarriers, a key advantage of the approach, demands perfect frequency and timing synchronization of the multiple users to within the duration of the cyclic prefix (CP). Since OFDM system shapes each of the subcarriers with a rectangular pulse, it suffers from high inter-carrier-interference in the presence of synchronization errors, especially in the uplink of mobile environments where it is challenging to track the Doppler shifts of different users [1]. Recently, filter bank based multicarrier (FBMC) system with offset quadrature amplitude modulation (OQAM), which detaches the above-mentioned problems, has gained significant attraction as a promising technique for replacing OFDM in future wireless communication systems [2], [3]. Utilizing well localized prototype filters, FBMC-OQAM systems alleviate the uplink synchronization requirements [4] and also, unlike OFDM, avoid CP required to combat the inter-symbol-interference [4], [5]. Similar to OFDM [6], [7], FBMC-OQAM can be united with multiple-input multiple-output (MIMO) systems to enhance the capacity through spatial multiplexing and/ or to decrease the bit error rate (BER)

through diversity [8]–[10]. For brevity, FBMC-OQAM is simply referred to as FBMC in the sequel.

The channel state information (CSI) at the receiver is mandatory for reliable data detection in coherent wireless communication systems. Hence channel needs to be estimated accurately at the receiver. In contrast to OFDM, the subcarrier orthogonality in FBMC systems holds in the real field only [5]. Consequently, a particular problem of *intrinsic interference* needs to be tackled, which renders channel estimation challenging in FBMC systems. Further, channel estimation in MIMO-FBMC systems requires the placing of zero symbols between adjacent training symbols to avoid inter-symbol-interference (ISI) due to the overlapping nature of the time domain FBMC symbols. This, in turn, requires careful examination of the intrinsic interference at the receiver to evaluate the resulting virtual training symbols and for construction of the orthogonal virtual training symbol matrix. However, despite this impediment, several schemes have been developed for channel estimation in FBMC systems. For example, reference [11] has designed an interference approximation method (IAM)-based channel frequency response (CFR) estimation scheme for FBMC systems. The IAM-based channel estimation scheme, which has gained significant popularity, is based on the commonly used assumption that the symbol time is sufficiently longer than the maximum channel delay spread, and also exploits the fact that each symbol in the FBMC system interferes with the symbols in its frequency-time (FT) neighborhood, over which the CFR can be assumed to be constant. The work in [12] extends the concept of IAM-based CFR estimation to MIMO-FBMC systems, and proposes a training-based least squares (LS) estimator. Authors [13] have investigated an IAM-based linear minimum mean square error (LMMSE) estimator for the joint noise variance and CFR estimation in distributed MIMO-FBMC downlink scenario with perfect knowledge of the channel covariance matrix. A scattered-pilot-based CFR estimation algorithm for MIMO-FBMC systems has been proposed in [14]. Reference [15] describes a scheme for training sequence design using zero-correlation zone sequences for IAM based channel estimation in MIMO-FBMC systems. A review of IAM-based CFR estimation for SISO and MIMO-FBMC systems has been presented in [16].

This paper first investigates a biased LS (BLS) MIMO-FBMC channel estimator by optimally scaling LS scheme,

and assuming a priori knowledge of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$, where $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$ denotes the trace of the channel correlation matrix $\mathbf{R}_{\mathbf{H}_m}$. In practice, if knowledge of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$ is not available at the receiver, an LS-BLS scheme is proposed, which utilizes LS-based estimate of the quantity $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$. Next, an LMMSE scheme for MIMO-FBMC channel estimation is examined by utilizing second-order statistical characteristics of intrinsic interference, and assuming a priori knowledge of the channel correlation matrix $\mathbf{R}_{\mathbf{H}_m}$. Further, a relaxed LMMSE (RLMMSE) scheme, which is less restrictive than the LMMSE technique, is proposed by assuming prior knowledge of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$. For the scenarios where $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$ is not known to the receiver, an LS-RLMMSE technique by using LS-based estimate of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_m}\}$ is investigated. Employing the second-order statistical properties of the intrinsic interference in FBMC systems, the mean square error (MSE) is derived for the aforementioned MIMO-FBMC channel estimators. It is shown analytically and numerically that the proposed estimators significantly outperform the estimator presented in [12]. The trade-off between the performance of the proposed estimators and the required amount of channel information is also numerically demonstrated.

The remainder of this paper is organized as follows. Next section details the MIMO-FBMC system model. Section-III presents the framework for training-based channel estimation in MIMO-FBMC systems. The LS-based and LMMSE-based proposed schemes for MIMO-FBMC channel estimation have been discussed in Subsections-III-A and III-C, respectively. Section-IV demonstrates numerical results and Section-V concludes the paper.

Notation: Upper and lower case bold face letters \mathbf{A} and \mathbf{a} denote matrices and vectors. The superscripts $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ represent the complex conjugate, transpose and Hermitian operators, respectively. The operator $\mathbb{E}[\cdot]$ denotes the expectation and $\text{Tr}\{\mathbf{A}\}$ represents the trace of a matrix \mathbf{A} . The operations $\Re\{\cdot\}$ and $\Im\{\cdot\}$ represent the real and imaginary parts, respectively, and $j \triangleq \sqrt{-1}$. The operation $\|\mathbf{A}\|_F^2$ denotes Frobenius norm of the matrix \mathbf{A} and \mathbf{I}_N represents an $N \times N$ identity matrix.

II. MIMO-FBMC SYSTEM MODEL

We consider a MIMO-FBMC system with N_t transmit antennas, N_r receive antennas and N subcarriers. Let the quantity $d_{m,n}^t$ at subcarrier index m and symbol time index n denotes a real OQAM symbol of duration $T/2$ for the t^{th} transmit antenna, which is drawn by extracting the real and imaginary parts of a T duration complex QAM symbol according to the rules given in [5]. The OQAM symbols are assumed to be spatially and temporally independent and identically distributed (i.i.d) with power P_d/N_t such that $\mathbb{E}[d_{m,n}^t (d_{m,n}^t)^*] = P_d/N_t$. The equivalent discrete-time baseband signal $s^t[k]$ at the t^{th} transmit antenna of the MIMO-FBMC system is expressed as [5]

$$s^t[k] = \sum_{m=0}^{N-1} \sum_{n \in \mathbb{Z}} d_{m,n}^t \chi_{m,n}[k], \quad \text{for } 1 \leq t \leq N_t, \quad (1)$$

where k denotes the sample index and the basis function $\chi_{m,n}[k]$ is defined as

$$\chi_{m,n}[k] = p[k - nN/2] \exp\{j2\pi mk/N\} e^{j\phi_{m,n}}. \quad (2)$$

The phase factor $\phi_{m,n}$ above is defined as $\phi_{m,n} = \frac{\pi}{2}(m+n) - \pi mn$ [5]. The symmetric real-valued pulse $p[k]$ of length L_p represents the impulse response of the discrete-time prototype filter of the FBMC system. To recover the real OQAM symbols $d_{m,n}^t$, the basis functions $\chi_{m,n}[k]$ satisfy the following real field orthogonality condition

$$\Re \left\{ \sum_{k=-\infty}^{+\infty} \chi_{m,n}[k] \chi_{\bar{m},\bar{n}}^*[k] \right\} = \delta_{m,\bar{m}} \delta_{n,\bar{n}}, \quad (3)$$

where $\delta_{m,\bar{m}}$ denotes the Kronecker delta with $\delta_{m,\bar{m}} = 1$ if $m = \bar{m}$ and zero otherwise. Let the quantity $\xi_{m,n}^{\bar{m},\bar{n}}$ be defined as $\xi_{m,n}^{\bar{m},\bar{n}} = \sum_{k=-\infty}^{+\infty} \chi_{m,n}[k] \chi_{\bar{m},\bar{n}}^*[k]$. Thus, we have

$$\xi_{m,n}^{\bar{m},\bar{n}} = \begin{cases} 1, & \text{if } (m, n) = (\bar{m}, \bar{n}) \\ j \langle \xi \rangle_{m,n}^{\bar{m},\bar{n}}, & \text{if } (m, n) \neq (\bar{m}, \bar{n}), \end{cases} \quad (4)$$

where $\langle \xi \rangle_{m,n}^{\bar{m},\bar{n}} = \Im\{\sum_{k=-\infty}^{+\infty} \chi_{m,n}[k] \chi_{\bar{m},\bar{n}}^*[k]\}$ denotes the imaginary part of the cross correlation between the basis functions [11]. Let $h^{r,t}[k]$, for $0 \leq k \leq L_h - 1$, denote an L_h tap multipath fading channel between the t^{th} transmit and r^{th} receive antennas. The signal received at the r^{th} receive antenna of the MIMO-FBMC system is given as

$$y^r[k] = \sum_{t=1}^{N_t} \left(s^t[k] * h^{r,t}[k] \right) + \eta^r[k], \quad (5)$$

where $1 \leq r \leq N_r$. The quantity $\eta^r[k]$ denotes the zero mean additive white Gaussian noise at the r^{th} receive antenna with power σ_η^2 , and is assumed to be known a priori. The demodulated signal on the r^{th} receive antenna at subcarrier index \bar{m} and symbol time index \bar{n} is obtained as

$$y_{\bar{m},\bar{n}}^r = \sum_{k=-\infty}^{+\infty} y^r[k] \chi_{\bar{m},\bar{n}}^*[k]. \quad (6)$$

Substituting the expressions for $s^t[k]$, $\chi_{\bar{m},\bar{n}}^*[k]$ and $y^r[k]$ from (1), (2) and (5), respectively in (6), and assume that the symbol time is sufficiently longer that the maximum channel delay spread [9], [16], the demodulated signal in (6) is simplified as [16, Eq. (9)], [9, Eq. (11)]

$$y_{\bar{m},\bar{n}}^r \approx \sum_{t=1}^{N_t} H_{\bar{m}}^{r,t} b_{\bar{m},\bar{n}}^t + \eta_{\bar{m},\bar{n}}^r, \quad \text{for } 1 \leq r \leq N_r. \quad (7)$$

Here $H_{\bar{m}}^{r,t}$ denotes the CFR from the t^{th} transmit to the r^{th} receive antenna at the \bar{m}^{th} subcarrier. The demodulated noise $\eta_{\bar{m},\bar{n}}^r = \sum_{k=-\infty}^{+\infty} \eta^r[k] \chi_{\bar{m},\bar{n}}^*[k]$ is distributed as $\mathcal{CN}(0, \sigma_\eta^2)$. The noise $\eta_{\bar{m},\bar{n}}^r$ is correlated across the subcarrier index \bar{m} and the symbol time index \bar{n} due to the real field orthogonality in FBMC systems. However, this correlation is negligible in practice due to well FT localized filters in FBMC system. Thus, similar to several works in the literature such as [10]–[12], [16], this correlation is neglected in our work. The term $b_{\bar{m},\bar{n}}^t = d_{\bar{m},\bar{n}}^t + j I_{\bar{m},\bar{n}}^t$ can

be considered as the virtual symbol given by the addition of the OQAM symbol $d_{\bar{m}, \bar{n}}^t$ and the *intrinsic interference* component $I_{\bar{m}, \bar{n}}^t$, which is expressed as

$$I_{\bar{m}, \bar{n}}^t = \sum_{(m, n) \in \Omega_{\bar{m}, \bar{n}}} d_{m, n}^t \langle \xi \rangle_{m, n}^{\bar{m}, \bar{n}}, \quad (8)$$

where $\Omega_{\bar{m}, \bar{n}}$ represents the neighborhood of the FT point (\bar{m}, \bar{n}) that does not include the point (\bar{m}, \bar{n}) . The term $I_{\bar{m}, \bar{n}}^t$ consists of ISI and inter-carrier-interference from the adjacent time-frequency symbols around the desired symbol at the FT index (\bar{m}, \bar{n}) . This is unlike OFDM systems wherein the former is suppressed using the CP, whereas the latter is removed using the complex field orthogonality among the subcarriers. For a well FT localized filter $p[k]$, a significant portion of the interference can be attributed to the first order neighborhood of the FT point (\bar{m}, \bar{n}) , denoted by $\Omega_{\bar{m}, \bar{n}} = \{(\bar{m} \pm 1, \bar{n} \pm 1), (\bar{m}, \bar{n} \pm 1), (\bar{m} \pm 1, \bar{n})\}$. Since the OQAM symbols are i.i.d. zero mean with power P_d/N_t , from (8), we have $\mathbb{E}[I_{\bar{m}, \bar{n}}^t] = 0$ and variance

$$\mathbb{E}[|I_{\bar{m}, \bar{n}}^t|^2] = \frac{P_d}{N_t} \sum_{(m, n) \in \Omega_{\bar{m}, \bar{n}}} |\langle \xi \rangle_{m, n}^{\bar{m}, \bar{n}}|^2. \quad (9)$$

Using (2) and (4), the term $\sum_{m=0}^{N-1} \sum_{n \in \mathbb{Z}} |\xi_{\bar{m}+m, \bar{n}+n}^{\bar{m}, \bar{n}}|^2$ for an $\alpha_0 \in \mathbb{Z}$, can be evaluated as

$$\begin{aligned} \sum_{m=0}^{N-1} \sum_{n \in \mathbb{Z}} |\xi_{\bar{m}+m, \bar{n}+n}^{\bar{m}, \bar{n}}|^2 &= N \sum_{k=-\infty}^{+\infty} \sum_{\alpha_0 \in \mathbb{Z}} p[k] p[k - \alpha_0 N] \\ &\times \sum_{n \in \mathbb{Z}} p[k - nN/2] p[k - (n + 2\alpha_0)N/2]. \end{aligned} \quad (10)$$

Since the prototype pulse $p[k]$ is symmetrical, for $\alpha_0 \neq 0$, it follows that for all k the summation $\sum_{n \in \mathbb{Z}} p[k - nN/2] p[k - (n + 2\alpha_0)N/2] = 0$, and for $\alpha_0 = 0$, we get $\sum_{n \in \mathbb{Z}} p^2[k - nN/2] = 2/N$ for all k [5, Eq. (81)]. Thus

$$\sum_{m=0}^{N-1} \sum_{n \in \mathbb{Z}} |\xi_{\bar{m}+m, \bar{n}+n}^{\bar{m}, \bar{n}}|^2 = \frac{2N}{N} \sum_{k=-\infty}^{+\infty} p^2[k] = 2, \quad (11)$$

where we have also used the fact that the pulse $p[k]$ has unity energy, i.e., $\sum_{k=-\infty}^{+\infty} p^2[k] = 1$. Since FBMC system comprises of well FT localized pulse shapes, we have

$$\begin{aligned} \sum_{(m, n) \in \Omega_{\bar{m}, \bar{n}}} |\langle \xi \rangle_{m, n}^{\bar{m}, \bar{n}}|^2 &\approx \sum_{m=0}^{N-1} \sum_{n \in \mathbb{Z}} |\xi_{\bar{m}+m, \bar{n}+n}^{\bar{m}, \bar{n}}|^2 - |\xi_{\bar{m}, \bar{n}}^{\bar{m}, \bar{n}}|^2 \\ &= 1. \end{aligned} \quad (12)$$

Thus, the variance of the intrinsic interference is

$$\mathbb{E}[|I_{\bar{m}, \bar{n}}^t|^2] = \frac{P_d}{N_t} \sum_{(m, n) \in \Omega_{\bar{m}, \bar{n}}} |\langle \xi \rangle_{m, n}^{\bar{m}, \bar{n}}|^2 \approx \frac{P_d}{N_t}. \quad (13)$$

Using the above result, the variance of the virtual symbol $b_{\bar{m}, \bar{n}}^t = d_{\bar{m}, \bar{n}}^t + j I_{\bar{m}, \bar{n}}^t$ can now be computed as

$$\mathbb{E}[|b_{\bar{m}, \bar{n}}^t|^2] = E[|d_{\bar{m}, \bar{n}}^t|^2] + E[|I_{\bar{m}, \bar{n}}^t|^2] \approx 2P_d/N_t. \quad (14)$$

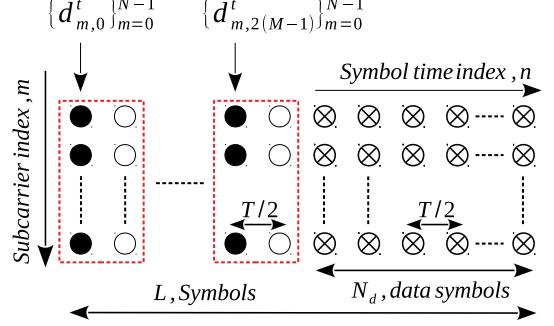


Figure 1. Frame structure for the t^{th} transmit antenna. The symbols, ●, ○ and ⊗ represent the training, zero and data symbols, respectively.

For analytical convenience, (7) can be succinctly formulated in vector form as

$$\mathbf{y}_{\bar{m}, \bar{n}} = \mathbf{H}_{\bar{m}} \mathbf{b}_{\bar{m}, \bar{n}} + \boldsymbol{\eta}_{\bar{m}, \bar{n}}, \quad (15)$$

where $\mathbf{y}_{\bar{m}, \bar{n}} = [y_{\bar{m}, \bar{n}}^1, y_{\bar{m}, \bar{n}}^2, \dots, y_{\bar{m}, \bar{n}}^{N_r}]^T \in \mathbb{C}^{N_r \times 1}$ is the vector of received symbols, $\boldsymbol{\eta}_{\bar{m}, \bar{n}} = [\eta_{\bar{m}, \bar{n}}^1, \eta_{\bar{m}, \bar{n}}^2, \dots, \eta_{\bar{m}, \bar{n}}^{N_r}]^T \in \mathbb{C}^{N_r \times 1}$ is the noise vector. The vector $\mathbf{b}_{\bar{m}, \bar{n}} = [b_{\bar{m}, \bar{n}}^1, b_{\bar{m}, \bar{n}}^2, \dots, b_{\bar{m}, \bar{n}}^{N_t}]^T \in \mathbb{C}^{N_t \times 1}$ comprises of the virtual symbols and $\mathbf{H}_{\bar{m}} \in \mathbb{C}^{N_r \times N_t}$ is the MIMO CFR matrix at the \bar{m}^{th} subcarrier whose element for the r^{th} row and the t^{th} column is given by $H_{\bar{m}}^{r,t}$. Using the second-order statistical properties of the intrinsic interference from (13), the covariance matrix of the virtual symbol vector in (15) is $\mathbb{E}[\mathbf{b}_{\bar{m}, \bar{n}} \mathbf{b}_{\bar{m}, \bar{n}}^H] \approx (2P_d/N_t) \mathbf{I}_{N_t}$.

III. TRAINING-BASED CHANNEL ESTIMATION FOR MIMO-FBMC SYSTEMS

Consider L OQAM symbols transmitted on each antenna and each subcarrier as per the frame structure illustrated in Fig. 1 for the t^{th} transmit antenna. Each frame comprises of M training symbols to be employed for channel estimation, followed by N_d data-bearing. In contrary to OFDM, the adjacent FBMC symbols interfere with each other in time domain due to the overlapping nature of the pulse-shaping filters. Thus, a zero symbol is typically inserted between adjacent training symbols to curb ISI as shown in Fig. 1 [11], [16]. In view of the inter-frame time gap commonly used in wireless communication, insertion of a zero symbol in the beginning of the frame is in general unnecessary [16]. The MIMO-FBMC pilot sequences with guard (zero) symbols require $2M$ OQAM symbols on each subcarrier, which is equivalent to M complex QAM symbols [16]. Hence, the training overhead required for channel estimation in MIMO-FBMC is similar to that of MIMO-OFDM, and does not incur any additional loss in spectral efficiency.

Evaluating (15) at the training symbol locations $n = 2i$ for $0 \leq i \leq M-1$ and stacking the outputs, one obtains

$$\mathbf{Y}_{\bar{m}} = \mathbf{H}_{\bar{m}} \mathbf{B}_{\bar{m}} + \boldsymbol{\eta}_{\bar{m}}, \quad (16)$$

where $\mathbf{Y}_{\bar{m}} = [\mathbf{y}_{\bar{m}, 0}, \mathbf{y}_{\bar{m}, 2}, \dots, \mathbf{y}_{\bar{m}, 2(M-1)}] \in \mathbb{C}^{N_r \times M}$ is the matrix of concatenated receive vectors and $\boldsymbol{\eta}_{\bar{m}} = [\boldsymbol{\eta}_{\bar{m}, 0}, \boldsymbol{\eta}_{\bar{m}, 2}, \dots, \boldsymbol{\eta}_{\bar{m}, 2(M-1)}] \in \mathbb{C}^{N_r \times M}$ is

the corresponding noise matrix. The matrix $\mathbf{B}_{\bar{m}} = [\mathbf{b}_{\bar{m},0}, \mathbf{b}_{\bar{m},2}, \dots, \mathbf{b}_{\bar{m},2(M-1)}] \in \mathbb{C}^{N_t \times M}$ is obtained by concatenation of the virtual training vectors. The t^{th} element of the virtual training vector $\mathbf{b}_{\bar{m},2i} \in \mathbb{C}^{N_t \times 1}$ at FT index $(\bar{m}, 2i)$ is given as $b_{\bar{m},2i}^t = d_{\bar{m},2i}^t + jI_{\bar{m},2i}^t$. For FBMC systems with a well FT localized pulse $p[k]$, the interference $I_{\bar{m},2i}^t$ for $0 \leq i \leq M - 1$ can readily be computed as

$$I_{\bar{m},2i}^t = \sum_{m \neq \bar{m}} d_{m,2i}^t \Im \left\{ \sum_{l=-\infty}^{+\infty} p^2[l] e^{j(\phi_{m,0} - \phi_{\bar{m},0})} e^{j2\pi(m-\bar{m})l/N} \right\} = \sum_{m \neq \bar{m}} d_{m,2i}^t \langle \xi \rangle_{m,0}^{\bar{m},0}. \quad (17)$$

Similar to [17], the training symbols herein are generated by extracting the real and imaginary parts of the random QAM symbol. Thus, for an orthogonal¹ training matrix $\mathbf{B}_{\bar{m}}$ [16], we have $\mathbf{B}_{\bar{m}}\mathbf{B}_{\bar{m}}^H = (2P_d/N_t)\mathbf{I}_M$. In the sequel, we consider $M = N_t$, which implies that each antenna transmits minimum number of training symbols to estimate the MIMO-FBMC channel.

A. LS MIMO-FBMC Channel Estimation

From (16), the training-based LS estimate of the CFR matrix $\mathbf{H}_{\bar{m}}$ is obtained as [16]

$$\hat{\mathbf{H}}_{\text{LS},\bar{m}} = \mathbf{Y}_{\bar{m}}\mathbf{B}_{\bar{m}}^\dagger = \mathbf{H}_{\bar{m}} + \boldsymbol{\eta}_{\bar{m}}\mathbf{B}_{\bar{m}}^\dagger, \quad (18)$$

where $\mathbf{B}_{\bar{m}}^\dagger = \mathbf{B}_{\bar{m}}^H(\mathbf{B}_{\bar{m}}\mathbf{B}_{\bar{m}}^H)^{-1}$ denotes the pseudo-inverse of the virtual training matrix $\mathbf{B}_{\bar{m}}$. The MSE of the LS estimator can be evaluated as

$$\begin{aligned} \Psi_{\text{LS}} &= \mathbb{E} \left\{ \left\| \mathbf{H}_{\bar{m}} - \hat{\mathbf{H}}_{\text{LS},\bar{m}} \right\|_F^2 \right\} = \mathbb{E} \left\{ \left\| \boldsymbol{\eta}_{\bar{m}}\mathbf{B}_{\bar{m}}^\dagger \right\|_F^2 \right\} \\ &= \text{Tr} \left\{ \mathbb{E} \left[(\boldsymbol{\eta}_{\bar{m}}\mathbf{B}_{\bar{m}}^\dagger)^H \boldsymbol{\eta}_{\bar{m}}\mathbf{B}_{\bar{m}}^\dagger \right] \right\} \\ &= \sigma_\eta^2 N_r \text{Tr} \left\{ (\mathbf{B}_{\bar{m}}\mathbf{B}_{\bar{m}}^H)^{-1} \right\}. \end{aligned} \quad (19)$$

B. BLS MIMO-FBMC Channel Estimation

The LS estimator in (18) does not achieve lowest MSE always [18]. To further reduce the MSE of the LS estimator, we propose a BLS estimator by introducing bias in the LS estimate [19]. Thus, MSE of this estimator is expressed as

$$\mathbb{E} \left\{ \left\| \mathbf{H}_{\bar{m}} - \epsilon \hat{\mathbf{H}}_{\text{LS},\bar{m}} \right\|_F^2 \right\}, \quad (20)$$

where the parameter ϵ is adjusted to minimise the MSE. Let the matrix $\mathbf{R}_{\mathbf{H}_{\bar{m}}} = \mathbb{E}[\mathbf{H}_{\bar{m}}^H\mathbf{H}_{\bar{m}}]$ comprises of the channel correlations. Then, using (18), (19) and the independence between the zero mean noise matrix $\boldsymbol{\eta}_{\bar{m}}$ and the CFR matrix $\mathbf{H}_{\bar{m}}$, the MSE of the BLS estimator can be simplified as

$$\mathbb{E} \left\{ \left\| \mathbf{H}_{\bar{m}} - \epsilon \hat{\mathbf{H}}_{\text{LS},\bar{m}} \right\|_F^2 \right\} = (1 - \epsilon)^2 \text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \} + \epsilon^2 \Psi_{\text{LS}}.$$

¹Let us consider $M = N_t = 2$. Assuming the first antenna transmits the training symbol $d_{\bar{m}}$ at symbol time indices 0 and 2 on the m^{th} subcarrier, the second antenna on the other hand uses the same preamble, but with reversed signs at the index 2. Thus, we have $b_{\bar{m},0}^1 = b_{\bar{m},2}^1 = b_{\bar{m},0}^2 = -b_{\bar{m},2}^2 = b_{\bar{m}}$. Therefore, $\mathbf{B}_{\bar{m}} = b_{\bar{m}}\mathcal{H}_2$ is an orthogonal virtual training matrix since \mathcal{H}_2 is the second-order Hadamard matrix.

The optimum value of the parameter ϵ can be computed as

$$\epsilon_o = \frac{\text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}}{\Psi_{\text{LS}} + \text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}}. \quad (21)$$

The corresponding minimized MSE of the BLS estimator can be determined by substituting ϵ_o in (20) as

$$\Psi_{\text{BLS}} = \frac{\Psi_{\text{LS}} \text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}}{\Psi_{\text{LS}} + \text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}}. \quad (22)$$

Above equation manifests that the MSE of the BLS estimator is always less than the MSE of the LS estimator, i.e. $\Psi_{\text{BLS}} < \Psi_{\text{LS}}$. The difference between these errors is significant in low signal to noise ratio (SNR) regime, where $\text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \} \ll \Psi_{\text{LS}}$. The estimate of the CFR matrix $\mathbf{H}_{\bar{m}}$ using the BLS estimator can now be obtained as

$$\hat{\mathbf{H}}_{\text{BLS},\bar{m}} = \epsilon_o \hat{\mathbf{H}}_{\text{LS},\bar{m}}. \quad (23)$$

The BLS estimator above requires a priori knowledge of $\text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}$. In practice, the requirement of $\text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}$ can be avoided by its LS-based estimate as $\text{Tr} \{ \hat{\mathbf{R}}_{\mathbf{H}_{\bar{m}}} \} = \text{Tr} \{ \hat{\mathbf{H}}_{\text{LS},\bar{m}}^H \hat{\mathbf{H}}_{\text{LS},\bar{m}} \}$. The estimator, which uses $\text{Tr} \{ \hat{\mathbf{R}}_{\mathbf{H}_{\bar{m}}} \}$ instead of $\text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \}$ is termed as the LS-BLS estimator.

C. LMMSE MIMO-FBMC Channel Estimation

From (16), the LMMSE estimate of the channel matrix $\mathbf{H}_{\bar{m}}$ is computed as

$$\hat{\mathbf{H}}_{\bar{m}} = \mathbf{Y}_{\bar{m}} \tilde{\mathbf{J}}_{\bar{m}}. \quad (24)$$

The matrix $\tilde{\mathbf{J}}_{\bar{m}}$, which minimizes MSE of the LMMSE estimator, is computed as follows

$$\tilde{\mathbf{J}}_{\bar{m}} = \arg \min_{\mathbf{J}_{\bar{m}}} \mathbb{E} \left\{ \left\| \mathbf{H}_{\bar{m}} - \mathbf{Y}_{\bar{m}} \mathbf{J}_{\bar{m}} \right\|_F^2 \right\}.$$

Let $e = \mathbb{E} \left\{ \left\| \mathbf{G}_{\bar{m}} - \mathbf{Y}_{\bar{m}} \mathbf{J}_{\bar{m}} \right\|_F^2 \right\}$ be the estimation error. Employing (16), the error e can readily be recast as

$$\begin{aligned} e &= \text{Tr} \{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \} + \text{Tr} \left\{ \mathbf{J}_{\bar{m}}^H (\mathbf{B}_{\bar{m}}^H \mathbf{R}_{\mathbf{H}_{\bar{m}}} \mathbf{B}_{\bar{m}} + \sigma_\eta^2 N_r \mathbf{I}_M) \mathbf{J}_{\bar{m}} \right\} \\ &\quad - \text{Tr} \left\{ \mathbf{J}_{\bar{m}}^H \mathbf{B}_{\bar{m}}^H \mathbf{R}_{\mathbf{H}_{\bar{m}}} \right\} - \text{Tr} \left\{ \mathbf{R}_{\mathbf{H}_{\bar{m}}} \mathbf{B}_{\bar{m}} \mathbf{J}_{\bar{m}} \right\}. \end{aligned} \quad (25)$$

The optimum matrix $\tilde{\mathbf{J}}_{\bar{m}}$ can now be computed by setting $\partial e / \partial \mathbf{J}_{\bar{m}} = 0$, which is given as

$$\tilde{\mathbf{J}}_{\bar{m}} = (\mathbf{B}_{\bar{m}}^H \mathbf{R}_{\mathbf{H}_{\bar{m}}} \mathbf{B}_{\bar{m}} + \sigma_\eta^2 N_r \mathbf{I}_M)^{-1} \mathbf{B}_{\bar{m}}^H \mathbf{R}_{\mathbf{H}_{\bar{m}}}.$$

Thus, the LMMSE estimate of the CFR matrix $\mathbf{H}_{\bar{m}}$ is determined as

$$\begin{aligned} \hat{\mathbf{H}}_{\text{LMMSE},\bar{m}} &= \mathbf{Y}_{\bar{m}} (\mathbf{B}_{\bar{m}}^H \mathbf{R}_{\mathbf{H}_{\bar{m}}} \mathbf{B}_{\bar{m}} + \sigma_\eta^2 N_r \mathbf{I}_M)^{-1} \\ &\quad \times \mathbf{B}_{\bar{m}}^H \mathbf{R}_{\mathbf{H}_{\bar{m}}}. \end{aligned} \quad (26)$$

Let $\mathbf{E}_{\bar{m}} = \mathbf{H}_{\bar{m}} - \hat{\mathbf{H}}_{\bar{m}}$ be the LMMSE error matrix. Thus, the MSE of the LMMSE channel estimator is obtained as

$$\begin{aligned} \Psi_{\text{LMMSE}} &= \text{Tr} \{ \mathbf{R}_{\mathbf{E}_{\bar{m}}} \} = \text{Tr} \{ \mathbb{E} [\mathbf{E}_{\bar{m}} \mathbf{E}_{\bar{m}}^H] \} \\ &= \text{Tr} \left\{ \left(\mathbf{R}_{\mathbf{H}_{\bar{m}}}^{-1} + \frac{1}{\sigma_\eta^2 N_r} \mathbf{B}_{\bar{m}} \mathbf{B}_{\bar{m}}^H \right)^{-1} \right\}. \end{aligned} \quad (27)$$

D. RLMMSE MIMO-FBMC Channel Estimation

The LMMSE estimator in (26) assumes that the channel correlation matrix $\mathbf{R}_{\mathbf{H}_{\bar{m}}}$ is known, an assumption which need not hold in practice. Therefore, we propose an estimator, which relax this assumption by replacing $\mathbf{R}_{\mathbf{H}_{\bar{m}}} = \rho \mathbf{I}_{N_r}$ in (26) as follows.

$$\begin{aligned} \hat{\mathbf{H}}_{\bar{m}} &= \rho \mathbf{Y}_{\bar{m}} (\rho \mathbf{B}_{\bar{m}}^H \mathbf{B}_{\bar{m}} + \sigma_\eta^2 N_r \mathbf{I}_M)^{-1} \mathbf{B}_{\bar{m}}^H \\ &\stackrel{(a)}{=} \frac{\rho}{\sigma_\eta^2 N_r} \mathbf{Y}_{\bar{m}} \left(\mathbf{I}_M - \frac{\rho N_t}{2P_d \rho + \sigma_\eta^2 N_r N_t} \mathbf{B}_{\bar{m}}^H \mathbf{B}_{\bar{m}} \right)^{-1} \mathbf{B}_{\bar{m}}^H, \end{aligned} \quad (28)$$

where (a) follows from the matrix inversion lemma. The parameter ρ is computed to minimize the MSE. For an orthogonal virtual training matrix $\mathbf{B}_{\bar{m}}$, the MSE of the estimator in (28) can be obtained as

$$\begin{aligned} \mathbb{E}\{\|\mathbf{H}_{\bar{m}} - \hat{\mathbf{H}}_{\bar{m}}\|_F^2\} &= \left(\frac{4P_d^2 \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}}{N_t^2} + 2\sigma_\eta^2 N_r P_d \right) \\ &\quad \left(\frac{\rho N_t}{2P_d \rho + \sigma_\eta^2 N_r N_t} - \frac{N_t \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}}{2P_d \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} + \sigma_\eta^2 N_r N_t^2} \right) \\ &\quad + \frac{\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} \sigma_\eta^2 N_r N_t^2}{2P_d \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} + \sigma_\eta^2 N_r N_t^2}. \end{aligned} \quad (29)$$

By differentiating above equation with respect to ρ and equating to zero, the optimum value of the parameter ρ is obtained as

$$\rho_o = \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} / N_t. \quad (30)$$

Substituting ρ_o in (28), the estimate of the CFR matrix $\mathbf{H}_{\bar{m}}$ using the proposed RLMMSE estimator is obtained as

$$\hat{\mathbf{H}}_{\bar{m},\text{RLMMSE}} = \mathbf{Y}_{\bar{m}} \left(\mathbf{B}_{\bar{m}}^H \mathbf{B}_{\bar{m}} + \frac{\sigma_\eta^2 N_r N_t}{\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}} \mathbf{I}_M \right)^{-1} \mathbf{B}_{\bar{m}}^H. \quad (31)$$

Further, the MSE of the RLMMSE estimator can be determined by substituting ρ_o in (29) as

$$\Psi_{\text{RLMMSE}} = \frac{\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} \sigma_\eta^2 N_r N_t^2}{2P_d \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} + \sigma_\eta^2 N_r N_t^2}. \quad (32)$$

For an orthogonal virtual training matrix $\mathbf{B}_{\bar{m}}$, the MSE of the LS estimator from (19) is $\Psi_{\text{LS}} = \sigma_\eta^2 N_r N_t^2 / (2P_d)$. Thus, it follows from (32) that the RLMMSE estimator always performs better than the LS estimator since $\Psi_{\text{RLMMSE}} < \Psi_{\text{LS}}$. The performance gain of the former over the latter is significant in the low SNR regime, i.e. when $2P_d \text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\} \ll \sigma_\eta^2 N_r N_t^2$. Although, the RLMMSE estimator requires a priori knowledge of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$, however, it is less restrictive than that of knowing $\mathbf{R}_{\mathbf{H}_{\bar{m}}}$, which, in turn, is less restrictive than that of knowing $\mathbf{H}_{\bar{m}}$ itself. In practice, the RLMMSE estimator can avoid the requirement of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$ by its LS-based estimate as $\text{Tr}\{\hat{\mathbf{R}}_{\mathbf{H}_{\bar{m}}}\} = \text{Tr}\{\hat{\mathbf{H}}_{\bar{m},\text{LS}}^H \hat{\mathbf{H}}_{\bar{m},\text{LS}}\}$. The resulting estimator, which uses $\text{Tr}\{\hat{\mathbf{R}}_{\mathbf{H}_{\bar{m}}}\}$ instead of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$ is termed as the LS-RLMMSE estimator. For an orthogonal virtual training matrix $\mathbf{B}_{\bar{m}}$, it can be obtained from (31) as

$$\mathbf{H}_{\bar{m},\text{LS-RLMMSE}} = \frac{N_t \text{Tr}\{\mathbf{Y}_{\bar{m}} \mathbf{Y}_{\bar{m}}^H\}}{2P_d (\text{Tr}\{\mathbf{Y}_{\bar{m}} \mathbf{Y}_{\bar{m}}^H\} + \sigma_\eta^2 N_r N_t)} \mathbf{Y}_{\bar{m}} \mathbf{B}_{\bar{m}}^H.$$

IV. SIMULATION RESULTS

A MIMO-FBMC-OQAM system with $N_t = N_r = 2$ and $N = 128$ subcarriers is considered. The OQAM data and training symbols are generated by extracting the real and imaginary parts of random QPSK (4-QAM) symbols. The Rayleigh fading model type B for indoor test environments that has $L_h = 6$ taps, with the power profile (in dB): [0.0, -3.6, -7.2, -10.8, -18, -25.2] and delay profile (in ns): [0, 100, 200, 300, 500, 700] is considered to model the channel between each transmit and receive antenna pair. The SNR of operation on each subcarrier is defined as $2P_d/\sigma_\eta^2$. The channel correlation matrix $\mathbf{R}_{\mathbf{H}_{\bar{m}}}$, similar to [20], has the following structure

$$[\mathbf{R}_{\mathbf{H}_{\bar{m}}}]_{l,k} = N_r \alpha^{|l-k|}, \quad (33)$$

where the parameter $0 \leq \alpha < 1$. The normalized mean square error (NMSE) at the \bar{m} th subcarrier is computed as $\|\hat{\mathbf{H}}_{\bar{m},\text{A}} - \mathbf{H}_{\bar{m}}\|_F^2 / \|\mathbf{H}_{\bar{m}}\|_F^2$, where $\text{A} \in (\text{LS}, \text{BLS}, \text{LS-BLS}, \text{LMMSE}, \text{RLMMSE}, \text{LS-RLMMSE})$. An isotropic orthogonal transform algorithm (IOTA) pulse shaping filter of duration $4T$ is used for the simulations. The discrete time pulse shaping filter $p[k]$ is obtained by setting the sampling interval $T_s = T/N$.

Fig. 2 shows NMSE comparison of the LS and the proposed BLS and LS-BLS channel estimation schemes. It is observed that the BLS estimator performs better than the LS and LS-BLS estimators, because the former exploits a priori knowledge of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$. The performance gain of this estimator is significant in low SNR regime, because as shown in (22), the MSE of the BLS scheme is considerably lower than that of the LS in low SNR regime. The LS-BLS estimator, which uses LS-based estimate of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$, performs considerably better than the LS estimator. Thus, it presents a good trade-off between the achieved NMSE performance and required channel knowledge. In high SNR regime, the MSE of both the BLS and LS-BLS estimators progressively approaches to that of the LS estimator. Thus, their performances coincide with that of the LS estimator.

Fig. 3 shows NMSE comparison of all the channel estimation schemes in this work. The LMMSE estimator performs best because it has a priori knowledge of the channel correlation matrix $\mathbf{R}_{\mathbf{H}_{\bar{m}}}$. The proposed RLMMSE scheme on the other side, which is less restrictive than the LMMSE since it requires prior knowledge of $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$, performs close to that of the LMMSE estimator. Therefore, the proposed RLMMSE scheme presents a good trade-off between the achieved performance and the required channel knowledge. The LS-RMMSE scheme, which acquires $\text{Tr}\{\mathbf{R}_{\mathbf{H}_{\bar{m}}}\}$ by its LS-based estimation, experiences performance degradation in low SNR regime. However, it still significantly outperforms the LS estimator in low SNR range. It is also observed that the proposed RLMMSE and LS-RLMMSE estimators perform similar to that of the proposed BLS and LS-BLS estimators, respectively. This happens because, as shown in (32), the MSE of the RLMMSE schemes reduces to that of the BLS scheme for an orthogonal virtual training matrix $\mathbf{B}_{\bar{m}}$.

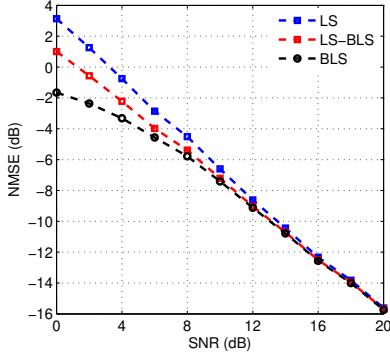


Figure 2. NMSE comparison of the proposed BLS and LS-BLS, and LS channel estimation schemes for MIMO-FBMC systems with orthogonal training, and $\alpha = 0.8$.

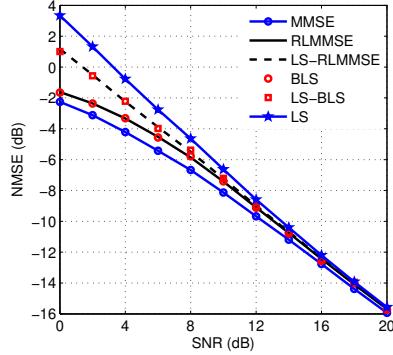


Figure 3. NMSE comparison of the LS, BLS, LS-BLS, LMMSE, RLMMSE and LS-RLMMSE channel estimation schemes for MIMO-FBMC systems with orthogonal training, and $\alpha = 0.8$.

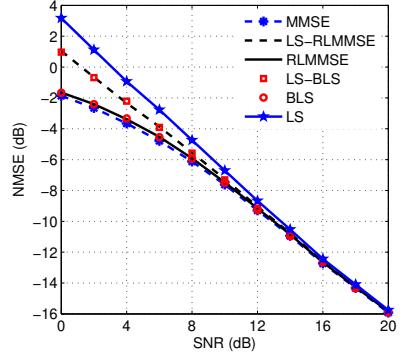


Figure 4. NMSE comparison of the LS, BLS, LS-BLS, LMMSE, RLMMSE and LS-RLMMSE channel estimation schemes for MIMO-FBMC systems with orthogonal training, and $\alpha = 0.5$.

Fig. 4 shows NMSE comparison of all the schemes for $\alpha = 0.5$ (weakly correlated channel). Since the LS, LS-BLS and LS-RLMMSE estimators do not use any prior knowledge of channel correlations, their NMSE performances are independent of α . The MMSE estimator, on the other hand, performs similar to that of the RLMMSE and BLS estimators, which is not surprising because the channel is weakly correlated.

V. CONCLUSIONS

The LS and linear MMSE schemes for MIMO-FBMC channel estimation have been investigated in the presence of limited and imperfect knowledge of the channel correlations. It has been observed that the performance of an estimator improves as the amount of a priori knowledge about channel correlations increases. It is demonstrated that the proposed schemes perform significantly better than the existing schemes in the low SNR regime. Future works can develop data-aided and sparse channel estimation schemes for MIMO-FBMC systems.

REFERENCES

- [1] M. Morelli, C. J. Kuo, and M. Pun, "Synchronization techniques for orthogonal frequency division multiple access (OFDMA): A tutorial review," *Proceedings of the IEEE*, vol. 95, pp. 1394–1427, 2007.
- [2] R. Nissel, S. Schwarz, and M. Rupp, "Filter bank multicarrier modulation schemes for future mobile communications," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 1768–1782, 2017.
- [3] M. Renfors, X. Mestre, E. Kofidis, and F. Bader, *Orthogonal Waveforms and Filter Banks for Future Communication Systems*. Academic Press, 2017.
- [4] A. Aminjavaheri, A. Farhang, A. RezazadehReyhani, and B. Farhang-Boroujeny, "Impact of timing and frequency offsets on multicarrier waveform candidates for 5g," *CoRR*, vol. abs/1505.00800, 2015.
- [5] P. Siohan, C. Siclet, and N. Lacaille, "Analysis and design of OFDM/OQAM systems based on filterbank theory," *IEEE Trans. Signal Processing*, vol. 50, no. 5, pp. 1170–1183, 2002.
- [6] K. Vasudevan, "Near capacity signaling over fading channels using coherent turbo coded OFDM and massive MIMO," *Int. Journal on Advances in Telecommunications*, vol. 10, pp. 22–37, 2017.
- [7] K. Vasudevan, "Turbo coded MIMO-OFDM," in *Twelfth International Conference on Wireless and Mobile Communications (ICWMC), Nov 13-17, 2016, Barcelona, Spain*, 2016, pp. 13–17.
- [8] A. I. Pérez-Neira, M. Caus, R. Zakaria, D. L. Ruyet, E. Kofidis, M. Haardt, X. Mestre, and Y. Cheng, "MIMO signal processing in offset-qam based filter bank multicarrier systems," *IEEE Trans. Signal Processing*, vol. 64, no. 21, pp. 5733–5762, 2016.
- [9] M. E. Tabach, J. Javaudin, and M. Hélard, "Spatial data multiplexing over OFDM/OQAM modulations," in *Proceedings of IEEE, ICC, Glasgow, Scotland, 24-28 June, 2007*, pp. 4201–4206.
- [10] P. Singh, R. Budhiraja, and K. Vasudevan, "SER analysis of MMSE combining for MIMO FBMC-OQAM systems with imperfect CSI," *IEEE Communications Letters*, pp. 1–1, 2018.
- [11] C. Lélé, J. Javaudin, R. Legouable, A. Skrzypczak, and P. Siohan, "Channel estimation methods for preamble-based OFDM/OQAM modulations," *European Transactions on Telecommunications*, vol. 19, no. 7, pp. 741–750, 2008.
- [12] E. Kofidis and D. Katsulis, "Preamble-based channel estimation in MIMO-OFDM/OQAM systems," in *2011 IEEE International Conf. on Signal and Image Processing Applications, ICSIPA, Kuala Lumpur, Malaysia, November 16-18, 2011*, pp. 579–584.
- [13] F. Rottenberg, Y. Medjahdi, E. Kofidis, and J. Louveaux, "Preamble-based channel estimation in asynchronous FBMC-OQAM distributed MIMO systems," in *Int. Symposium on Wireless Communication Systems (ISWCS), Brussels, August 25-28, 2015*, pp. 566–570.
- [14] J.-P. Javaudin and Y. Jiang, "Channel estimation in MIMO OFDM/OQAM," in *Workshop on Signal Processing Advances in Wireless Communications, SPAWC*. IEEE, 2008, pp. 266–270.
- [15] S. Hu, Z. L. Liu, Y. L. Guan, C. Jin, Y. Huang, and J. Wu, "Training sequence design for efficient channel estimation in MIMO-FBMC systems," *IEEE Access*, vol. 5, pp. 4747–4758, 2017.
- [16] E. Kofidis, D. Katsulis, A. A. Rontogiannis, and S. Theodoridis, "Preamble-based channel estimation in OFDM/OQAM systems: A review," *Signal Processing*, vol. 93, no. 7, pp. 2038–2054, 2013.
- [17] P. Singh, E. Sharma, K. Vasudevan, and R. Budhiraja, "CFO and channel estimation for frequency selective MIMO-FBMC/OQAM systems," *IEEE Wireless Communications Letters*, 2018.
- [18] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, "Linear minimax regret estimation of deterministic parameters with bounded data uncertainties," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2177–2188, 2004.
- [19] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, no. 3, pp. 497–508, 1973.
- [20] A. B. Gershman, C. F. Mecklenbräuker, and J. F. Böhme, "Matrix fitting approach to direction of arrival estimation with imperfect spatial coherence of wavefronts," *IEEE Trans. Signal Processing*, vol. 45, no. 7, pp. 1894–1899, 1997.

Hardware Implementation of Filtered OFDM for BB-PLC using Software Defined Radio

Sumesh K P*, Ankit Dubey†, and Trilochan Panigrahi‡

Department of ECE, National Institute of Technology Goa, Farmagudi, Ponda, Goa 403401, India

{*sumeshkp, †ankit.dubey, ‡tpanigrahi}@nitgoa.ac.in

Abstract—Among several in-home communication systems, broadband over power line communication (BB-PLC) provides better connectivity at cheaper installation cost without disturbing the existing infrastructure. Conventional broadband systems, including BB-PLC, use the orthogonal frequency division modulation (OFDM) to combat the issue of frequency selective fading at the cost of extra bandwidth. However, it is proposed and shown that the spectral efficiency of the OFDM can be improved either by using filter bank multi-carrier (FBMC) or filtered OFDM (F-OFDM). In the former, all the sub-carriers of an OFDM symbol are filtered, however, in the latter, only the final OFDM symbol is filtered. Thus, it is a cost-effective solution to use the F-OFDM over the FBMC. This paper presents the hardware implementation of the F-OFDM for BB-PLC. Different prototype low-pass filters are used for comparative analysis, especially for the BB-PLC. The simulations are carried out in National Instruments' LabVIEW software and the hardware implementation of the transceiver is done on National Instruments' Software Defined Radio (SDR) set called USRP 2920. Polycab 1.5 SQ mm home wiring power cable is used as the communication channel in testing the BB-PLC system. From the power spectral density analysis, it is concluded that the F-OFDM has better spectral efficiency than the conventional OFDM. Further, it is observed that as the length of the filter increases the better spectral efficiency is achieved but at the cost of increased complexity.

Index Terms—BB-PLC, FBMC, F-OFDM, OFDM, power line communication, USRP.

I. INTRODUCTION AND MOTIVATION

At present, the broadband internet technology is playing a prime role in the progress of human life and the betterment of the nation, as it bestows breakneck services like video streaming, e-mail, internet banking, e-commerce, stock market trading, etc. [1]. There are many techniques which are currently being used to provide broadband internet access, which comprises broadcast satellites, fixed wireless and digital subscriber line (DSL) and cable modem [2]. However, the capacity of these technologies to penetrate deeper into the rural areas of the country is poor. Therefore, the accessibility to the broadband internet is limited. Most of the places, whether it is a rural or urban area, would be having electrical/power line cables. These cables are deployed already and it penetrates to a higher extent into the rural areas. Thus using these cables for the purpose of data transmission would be much cheaper compared to the expensive equipment and cables for the DSL lines and cable modems. Also, the maintenance and repairing of the power line infrastructure is much simpler and cheaper compared

to other methods [3], [4]. In power line cables, the power transmission uses only 50-60 Hz spectrum, the rest of the spectrum is not utilized. Hence, the unused higher spectrum in the electrical cables is proposed to be utilized for the purpose of data transmission and in-general, it is termed as the power line communication (PLC) [4]. However, when using the same for the broadband access, it is called the broadband over PLC (BB-PLC) [4], [5].

The major challenge in a BB-PLC system arises during the low power data transmission alongside the high power electrical energy [6], [7]. The data transmission through the electrical cable suffers from the additive impulsive noise along with the background noise [8]. Moreover, while using the higher frequency band for the BB-PLC, the channel behaves as a frequency selective fading channel [9], [10]. Thus, the orthogonal frequency division modulation (OFDM) has been proposed for the BB-PLC system that is able to cope up with the bleak channel conditions [4], [8], [11]. Moreover, the OFDM is currently being used in LTE, Wi-Fi, Wi-MAX, etc. because of its ability to work in harsh channel environment without using complex equalization filters [12]. The OFDM is both; a modulation and a multiplexing technique. Despite having so many advantages, its spectral leakage in side-lobes is a challenging problem [13]. The side-lobe leakage causes inter symbol interference (ISI) between adjacent OFDM symbols. Thus, to avoid ISI a guard band has to be provided between the adjacent channels. This, in turn, reduces the bandwidth efficiency of the OFDM based system [14]. However, it is proposed and shown that the spectral efficiency of the OFDM can be improved either by using filter bank multi-carrier (FBMC) or filtered OFDM (F-OFDM) [15]. In the former, all the sub-carriers of an OFDM symbol are filtered, however, in the later, only the final OFDM symbol is filtered. Thus, F-OFDM offers a cost-effective and less complex solution as compared to that of the FBMC. In [16], a prototype filter-based approach is proposed to truncate the sideband of the OFDM symbols. However, up to the authors' knowledge, the hardware implementation of F-OFDM has not been reported.

The main objective of this paper is to describe the hardware implementation of the F-OFDM in real time conditions for BB-PLC in the Indian scenario. Moreover, the effect of different prototype filters and their length on the spectral efficiency are demonstrated through the numerical results using the measure-

ments obtained.

The rest of the paper is organized as below. BB-PLC is described in section II. Section III briefs out the PLC channel model and Section IV explains the F-OFDM transceiver created in LabVIEW. Numerical results and analysis are provided in the subsequent section. Finally, conclusions are drawn in Section VI.

II. BB-PLC

As mentioned earlier, the BB-PLC is a technology that provides broadband access using electrical cables [4], [5], [17]. As the electrical supply signals use only a few Hertz spectrum, the remaining wide spectrum can be used for communication purpose. It has also been shown theoretically that the broadband spectrum will not be affected by the frequency of the energy supply signals [18]. BB-PLC uses multi-carrier modulation (MCM) techniques to get high throughput. In an MCM technique, the data is modulated to adjacently spaced slowly varying multiple carriers instead of a fast varying single carrier [19]. The main advantages of an MCM system include immunity against the frequency selective fading, less susceptibility to interference caused by the impulse noise, and the ability to fight off inter symbol interference (ISI). Moreover, an MCM system provides maximum utilization of the available spectrum [20].

BB-PLC technology operates on the frequency range of 2 to 30 MHz, as of now. Exploring the spectrum beyond this range is a big challenge because these power cables are not designed for the communication purpose. Another reason is that the unavailability of the high-frequency sophisticated power-line couplers [21]. The time division duplexing (TDD) and frequency division duplexing (FDD) can be adopted for the BB-PLC. In the FDD, two different frequencies for up-link and down-link are used whereas, in the TDD same frequency band is utilized for up-link and down-link but at different time slots [4].

The main components of the BB-PLC are the head end, repeaters, and the customer premise equipment or a BB-PLC modem. The head end injects the broadband internet into the power cables. As the broadband signal progresses through the cable, it faces distortion. Hence, repeaters, which are basically amplify-and-forward or decode-and-forward type, are used to regenerate the degrading signal [22]–[24]. In the end, the BB-PLC modem is used to tap-out the broadband signals from the power-line cable [25].

III. PLC CHANNEL MODEL

A multipath channel model is used in the analysis. In BB-PLC, a transmitted signal reaches the receiver several times with different delays due to the multiple reflections suffered by the signal. Thus, the rudimentary behavior of the BB-PLC channel, where M reflections are potent, can be expressed by its impulse response as [4], [26]

$$h(t) = \sum_{i=0}^{M-1} g_i \delta(t - \tau_i), \quad (1)$$

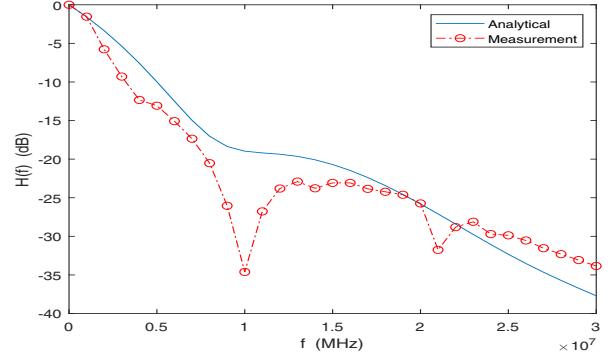


Fig. 1. Channel response: Analytical vs measurement

where the coefficient τ_i denotes the individual delays of an echo, and the coefficient g_i is the corresponding attenuation of a reflection. The Fourier transform of the above equation delivers the complex transfer function given as

$$H(f) = \sum_{i=0}^{M-1} g_i e^{-j2\pi f \tau_i}. \quad (2)$$

In real time scenario, τ_i and coefficients g_i depend on the cable length and frequency. Thus, g_i can be expressed as

$$g_i \Rightarrow g(f, l_i) = k_i e^{-\alpha(f)l_i}. \quad (3)$$

Adding up the consequences of length dependent attenuation, multipath propagation, and frequency dependent attenuation, the complete transfer function can be written as [4], [26]

$$\begin{aligned} H(f) &= \sum_{i=0}^{M-1} k_i e^{-\alpha(f)l_i} e^{-j2\pi f l_i / v_p} \\ &= \sum_{i=0}^{M-1} k_i e^{-[\alpha(f) + j\beta]l_i}, \end{aligned} \quad (4)$$

where v_p is the phase velocity, k_i is a constant, l_i is the length of the i^{th} channel, and $\alpha(f)$ is the frequency dependent attenuation coefficient, given as

$$\alpha(f) \approx c_1 \sqrt{f} + c_2 f \approx a_0 + a_1 f^{0.5...1}, \quad (5)$$

where a_0 and a_1 are constants that are provided in [4].

Thus, it can be observed that the BB-PLC channel is a periodically varying one. It has alternative peaks and notches that show frequency selective nature of the channel. It is important to select communication carrier frequency such that it should suffer minimum attenuation.

In fig 1, we have plotted the analytical and experimental channel measurements. For the analytical curve, the main channel length is taken as 10 m and the reflected path length as 18 m with $M = 2$. The coefficients, a_0, a_1, k_0 , and k_1 are taken as 0, 7.8×10^{-9} , 0.55, and 0.45, respectively. For experimental channel, we have taken a 10 m long polycab 1.5 SQ mm 1100 V cable with 25 strands which is commonly used for home wiring and the measurements are obtained using

TABLE I
FILTER COEFFICIENTS

Filter	Length	Coefficients
dB3	6	0.0352263, -0.0854413, -0.135011, 0.4598780, 0.8068920, 0.3326710
Symlet	6	0.0384600, -0.0769200, -0.461538, -0.461538, -0.07692, 0.03846
CDF 9/7	9	0.0267, -0.0169, -0.0782, 0.2669, 0.6029, 0.2669, -0.0782, -0.0169, 0.0267

a spectrum analyzer. The comparative analysis shows a little variation between the measurements and analytically obtained channels. However, one can observe that both of them have a notch at a frequency of 10 MHz.

IV. F-OFDM TRANSCEIVER

In this section, we discuss the hardware implementation of the F-OFDM in detail. We start with the OFDM block diagram followed by filter design and implementation of the F-OFDM.

A. OFDM

The OFDM transmitter consists blocks of data symbol mapper, inverse fast Fourier transform (IFFT), and cyclic prefix (CP). The data symbol mapper block takes a chunk of data and maps them to complex symbols (using digital modulation technique). The IFFT block uses orthogonal subcarriers to modulate each symbol. The CP block adds a part of the modulated symbols at the beginning of the OFDM symbol to combat the inter-symbol-interference (ISI). The reverse is performed at the receiver side. Mathematically, the OFDM symbol can be represented as [16],

$$s(n) = \sum_{i=0}^{L-1} s_l(n - l(N + N_g)), \quad (6)$$

here,

$$s_l(n) \triangleq \sum_{m=m'}^{m'+M-1} d_{l,m} e^{j2\pi mn/N}, -N_g < n < N, \quad (7)$$

where, N is the IFFT length, N_g is the cyclic prefix length, M represents the number of consecutive sub-carriers used to form a single OFDM symbol, and L represents the number of OFDM symbols. Further, $d_{l,m}$ denotes the data symbol on the m^{th} sub-carrier of the l^{th} OFDM symbol. Furthermore, $m', m' + 1, \dots, m' + M - 1$ denote the allocated sub-carrier range.

B. Filter Design

Several prototype filter designs are available in the literature [16]. In this paper, we have considered three different filters, namely; dB3, Symlet, and Cohen-Daubechies-Feauveau 9/7 (CDF 9/7), for the prototype filter design [27]–[29]. To reduce the complexity and the implementation cost, all the filters are

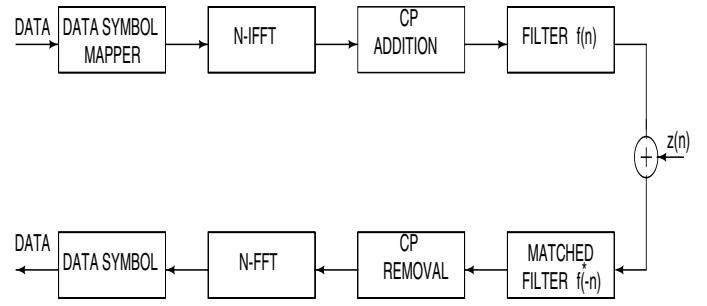


Fig. 2. Block Diagram of F-OFDM

designed with lengths lesser than that of filters proposed in [16]. Out of the three filters, dB3 and Symlet are of the same length and the third one is comparatively of higher length. A prototype design for the CDF 9/7 filter is given below [27]

$$P_k(z) = z^{-k} \left(\frac{1 - z^{-1}}{2} \right)^{2k} \sum_{n=0}^{k-1} 2 \frac{(k+n+1)}{n} (-z)^n \left(\frac{(1 - z^{-1})}{2} \right). \quad (8)$$

For $k = 4$, i.e. for the filter length 9, the coefficients are tabulated in the Table I. Similarly, filter lengths and corresponding coefficients for the rest of the filters are calculated using [28], [29] and values are tabulated in the Table I.

C. F-OFDM Implementation

Fig. 2 shows the block diagram of the F-OFDM system. It consists of a data symbol mapper, an N-IFFT block, a CP addition block, and a filter block at the transmitter side. At the receiver side, it consists a matched filter, CP removal block, an N-FFT block, and a data retriever block. Thus, the F-OFDM symbol is formed by passing the OFDM symbol through the prototype spectrum shaping filter. Hence, F-OFDM symbol can be expressed as a convolution of the OFDM symbol, $s(n)$, with the prototype filter, $f(n)$, as

$$s'(n) = s(n) * f(n) \quad (9)$$

The signal generated after this process would be frequency limited at both ends which will be transmitted after channel modulation. Different steps involved to implement F-OFDM are explained as follows.

1) *Symbol Mapper*: The symbol mapper consists of a quadrature amplitude modulation (QAM) modulator, data interleaver, and zero padding block. Initially, the file to be transmitted is fragmented in blocks of data and pilots (labels) are inserted for the purpose of reconstruction after reception. Then, the data is converted to a stream of binary digits. We use 4-QAM to convert the bits to complex symbols by processing two bits at a time. For more data rates, it can be easily modified to 8-QAM, 16-QAM or more. The data interleaving block adds 1 redundant symbol after every 5th data symbol coming from QAM modulator for synchronization purpose. After the interleaving, there will be 25 redundant data symbols in an OFDM symbol. After this, zero padding is performed at both

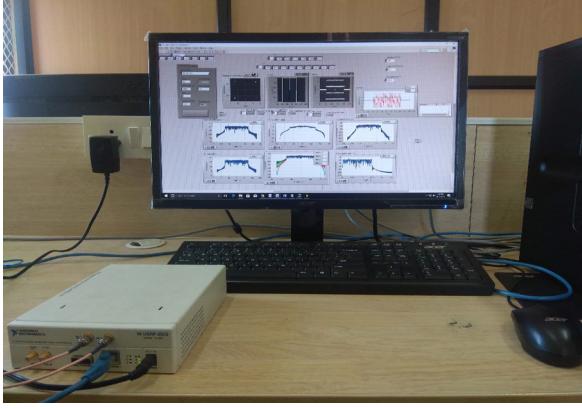


Fig. 3. Experimental Setup

ends of the data to increase the frequency resolution of the inverse discrete Fourier transform (IDFT). At both the ends, 53 zeros are added so that the entire data set length becomes 256.

2) *IFFT*: In the IFFT block, we take 256 point-IDFT to obtain a single OFDM symbol. In general, N different orthogonal carriers are modulated using N complex symbols to produce one OFDM symbol.

3) *CP Addition*: Next step of the transmitter is to add the CP to the data. CP is the tail end of the data stream which is prefixed to the data. Even though the CP is overhead, it is necessary for synchronization and to avoid the ISI. We use 64 data points as CP. Thus, the total OFDM symbol contains 320 data points.

4) *Filtering*: The final stage is to limit the spectrum of the generated OFDM symbol, which is achieved by passing the generated symbol through the prototype filter $f(n)$. In the end, like any other communication system, channel modulation is essential in this case. Thus, we carrier modulated each F-OFDM symbol using 50 MHz sine wave with IQ rate of 5 MHz.

At the receiver side, after channel demodulation, a matched filter $f^*(-n)$ is used to reconstruct the OFDM signal. This reconstructed symbol is then passed to the next stage for further processing. A Vande-beek detector is used to synchronize the signal with the transmitter side signal [30]. The Vande-beek algorithm detects the starting of an OFDM symbol using the CP contained in the symbol. Thereafter, the CP is removed and the symbol is passed through 256-point FFT that produces parallel complex data symbols. Now, we perform the zero padding removal and de-interleaving at this point. Using the padded zeros, channel estimation is carried out and equalization is performed in the following step. In the next step, the equalized complex symbols are converted back to data bits using QAM-demodulator. From the received data bits, the file is reconstructed in the final step.

D. Experimental Setup

The real-time F-OFDM transceiver for BB-PLC is implemented using National Instruments LabView and Universal

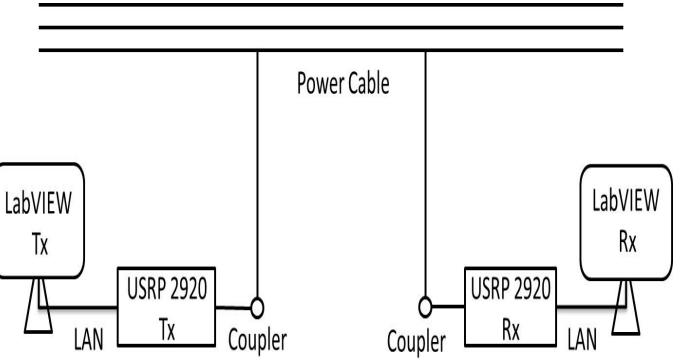


Fig. 4. Experimental Setup Block Diagram

Software Radio Peripheral-2920 (USRP-2920).

1) *USRP-2920*: The USRP-2920 is a tunable RF transceiver with a high-speed analog to digital converter and digital to analog converter for streaming base-band I and Q signals to a host PC over 1 Gigabit Ethernet. Its bandwidth capability is 20 MHz and it can handle frequencies from 50 MHz to 2.4 GHz. It can be used for the following applications also: white space; broadcast FM; public safety; land-mobile, low-power unlicensed devices on industrial, scientific, and medical (ISM) bands; sensor networks; cell phone; amateur radio; or GPS.

2) *The Experimental Setup*: The experimental setup is shown in Fig. 3 and its block diagram are shown in Fig. 4. A text file is taken for experiment and it is fragmented and sent through the home wiring (power) cable, after adding error correction codes and synchronization bits for the reconstruction purpose. The F-OFDM based receiver is kept in a separate room within a building and the transmitted file by the F-OFDM based transmitter is received successfully. The cable used for the experiment is a polycab 1.5 SQ mm 1100v with 25 strands which are commonly used for home wiring. The cable is connected to the USRP-2920 using crocodile cables at both the transmitter and receiver side. The USRPs are connected to the computer through a gigabit Ethernet cable. We have transmitted and received binary data bits through home wiring power line cable successfully using F-OFDM technique.

V. RESULTS AND ANALYSIS

In this section, we compare the F-OFDM with conventional OFDM (CP-OFDM) using the power spectral density (PSD). Further, we analyze the effect of filter length on the spectrum efficiency of the F-OFDM.

Fig. 5 shows the PSDs of both the MCM schemes, the F-OFDM and the conventional OFDM, with the same parameters as described in Section IV. In the F-OFDM, the low pass prototype filters are used to clip down the leakages in the side lobes. The CDF 9/7 filter is used as a prototype filter to obtain the PSD for the F-OFDM [27]. It has 9 filter coefficients at the transmitter side and at the receiver side, it has a filter of length 7 which is matched to the filter at the transmitter side.

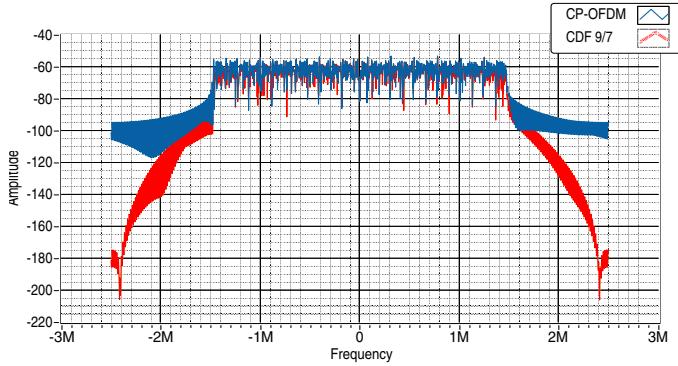


Fig. 5. CP-OFDM vs Filtered OFDM

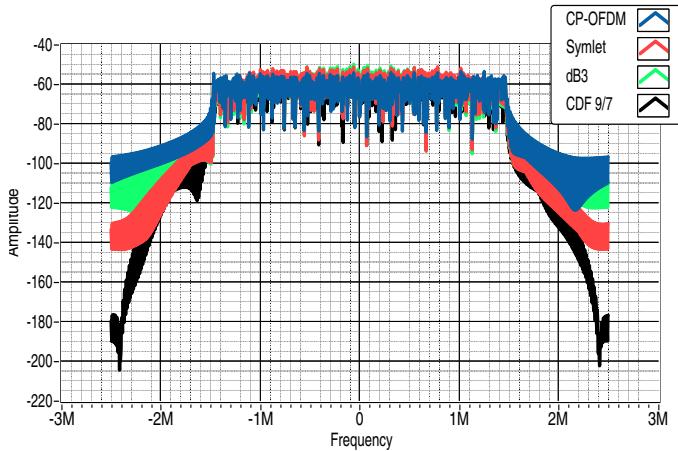


Fig. 6. PSD of Different Filters

We compare the spectrum efficiency of the two MCM schemes using percentage power reduction (PR) in the side lobes. The numerical results obtained from Fig. 5 shows that the PR in the side lobes of the F-OFDM scheme with CDF 9/7 filter is around 83.34%. That means that the side lobe leakage in the conventional OFDM is comparatively very high. Thus, it can be concluded that, more number of parallel channels can be accommodated in the available bandwidth while using F-OFDM.

TABLE II
SIDE LOBE POWER

Signal/Filter	Sidelobe Power (mW)	Percentage Reduction (%)
OFDM	6.50683	—
dB3	1.72461	73.49
Symlet	1.57481	75.79
CDF 9/7	1.08402	83.34

Fig. 6 shows the PSD of the signals when different filters are used. It comprises of PSD of conventional OFDM and F-OFDM with different prototype filters. We use dB3 filter [28], Symlet

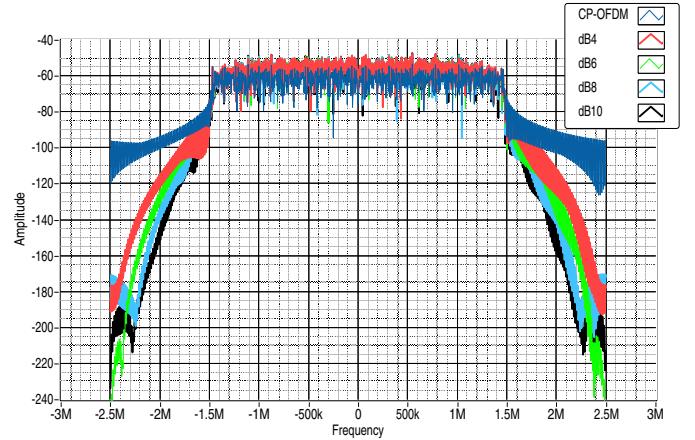


Fig. 7. CP-OFDM and dB filter with different length

filter, [29] and CDF 9/7 filters as prototype filters to obtain the PSDs for the F-OFDM. The dB3 filter and Symlet filters consist of 6 filter coefficients whereas the CDF filter consists of 9 filter coefficients. Comparing the equal length filters, the Symlet and the dB3, the Symlet outperforms the dB3 filter. Further, it can be observed that the CDF 9/7 perform superior to both the filters. Table II shows the power distribution of the side lobes of the corresponding spectrum. It can be observed that using F-OFDM side-lobe leakage can be reduced drastically. The percentage PR in side lobes when a dB3 filter is used is 73.49% compared to the conventional OFDM. Whereas, it is 75.79% for the Symlet filter. Moreover, if we use the CDF 9/7 filter, it provides 83.3% PR in side lobes against the conventional OFDM.

In the end, we compare the effect of filter length on spectral efficiency with the help of Fig. 7. We select a $dB\ell$ filter with different ℓ for the analysis. A $dB\ell$ filter has 2ℓ filter coefficients. We obtain PSDs of the F-OFDM with dB4, dB6, dB8, and dB10 filters. From the observations, it can be concluded that for the same type of filter, the spectral efficiency improves as the number of filter coefficients increases. However, it is also clear that the improvement is not linear.

VI. CONCLUSION

In this work, we have presented the hardware implementation of the F-OFDM for the BB-PLC and successfully exchanged files between two devices through electrical cables. The simulations are carried out in National Instruments' LabVIEW software and the hardware implementation of the transceiver is done on National Instruments' Software Defined Radio (SDR) set known as the USRP 2920. Polycab 1.5 SQ mm with 25 strands standard home wiring power cable is used as the communication channel in testing the BB-PLC system. We have observed that the F-OFDM can be used to achieve better spectral efficiency while enjoying the other benefits of conventional OFDM. We have also observed that the CDF 9/7 filter outperforms the dB3 and the Symlet filters as a candidate for the F-OFDM. Furthermore, it is observed that the more the

length of the filter the better the spectral efficiency. However, at the cost of increased complexity.

ACKNOWLEDGMENT

This work was supported in part by the Department of Science and Technology (DST), Govt. of India (Ref. No. TMD/CERI/BEE/2016/059(G)) and the Science and Engineering Research Board (SERB), Govt. of India through its Early Career Research (ECR) Award (Ref. No. ECR/2016/001377).

REFERENCES

- [1] P. K. Ray and A. Hazra, "Broadband powerline communication an Indian experience," in *Proc. IEEE International Symposium on Power Line Communications and its Applications (ISPLC)*, Udine, Italy, 2011, pp.364–369.
- [2] D. Bogojevic and N. Gospic, "Some aspects of broadband Internet growth and limits," in *Proc. 2011 10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS)*, Serbia, 2011, pp. 615–618.
- [3] S. Galli, A. Scaglione, and Z. Wang, "For the grid and through the grid: The role of power line communications in the smart grid," *Proceedings of the IEEE*, vol. 99, no. 6, pp. 998–1027, Jun. 2011.
- [4] H. C. Ferreira, L. Lampe, J. Newbury, and T. G. Swart, *Power Line Communications: Theory and Applications for Narrowband and Broadband Communications over Power Lines*, Singapore: Wiley, 2010.
- [5] L. R. Varshney, "Transporting information and energy simultaneously," in *Proc. 2008 IEEE International Symposium on Information Theory*, Toronto, ON, 2008, pp. 1612–1616.
- [6] S. Moya, M. Hadad, P. Donato, M. Funes, and D. Carrica, "Channel estimation and equalization of broadband PLC systems - Part 2: Comparison between Single Carrier and OFDM approaches," in *Proc. 2016 IEEE Biennial Congress of Argentina (ARGENCON)*, Buenos Aires, Argentina, 2016, pp. 1–6.
- [7] S. Liu, F. Yang, W. Ding, J. Song, and Z. Han, "Impulsive noise cancellation for MIMO-OFDM plc systems: A structured compressed sensing perspective," in *Proc. 2016 IEEE Global Communications Conference (GLOBECOM)*, Washington, DC, 2016, pp. 1–6.
- [8] Y. H. Ma, P. L. So, and E. Gunawan, "Performance analysis of OFDM systems for broadband power line communications under impulsive noise and multipath effects," *IEEE Trans. Power Deliv.*, vol. 20, no. 2, pp. 674–682, 2005.
- [9] A. Emleh, A. de Beer, H. Ferreira, and A. H. Vinck, "Interference detection on powerline communications channel when in-building wiring system acts as an antenna," in *Proc. ELMAR-2013*, Zadar, 2013, pp. 141–144.
- [10] A. Yousaf, F. Khan, Z. Hameed, and H. Ali, "Deployment of smart grid on narrowband power line communication using OFDMA," in *Proc. 2018 9th International Renewable Energy Congress (IREC)*, Hammamet, 2018, pp. 1–6.
- [11] Z. Ma, A. Gholamzadeh, B. Tang, S. Dang, and S. Yang, "MATLAB based simulation of the efficiency of the complex OFDM on power line communication technology," in *Proc. 2014 Fourth International Conference on Instrumentation and Measurement, Computer, Communication and Control*, Harbin, 2014, pp. 374–378.
- [12] T. Larhzaoui, F. Nouvel, J. Y. Baudais, P. Degauque, and V. Dgardin, "OFDM PLC transmission for aircraft flight control system," in *Proc. 18th IEEE International Symposium on Power Line Communications and Its Applications*, Glasgow, 2014, pp. 220–225.
- [13] Y. Zhang, L. Lin, and X. Ma, "A simple soft-in soft-out equalization for highly mobile OFDM systems with block Markov superposition transmission," in *Proc. 2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, Sydney, NSW, 2017, 1–5.
- [14] Y. Qiu, Z. Liu, and D. Qu, "Filtered bank based implementation for filtered OFDM," in *Proc. 2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC)*, Macau, 2017, pp. 15–18.
- [15] R. Zakaria and D. Le Ruyet, "Theoretical analysis of the power spectral density for FFT-FBMC signals," *IEEE Communications Letters*, vol. 20, no. 9, pp. 1748–1751, Sep. 2016.
- [16] J. Abdoli, M. Jia, and J. Ma, "Filtered OFDM: A new waveform for future wireless systems," in *Proc. 2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, 2015, pp. 66–70.
- [17] S. Galli and O. Logvinov, "Recent developments in the standardization of power line communications within the IEEE," *IEEE Communications Magazine*, vol. 46, no. 7, pp. 64–71, Jul. 2008.
- [18] S. W. Oh et al., "Cognitive power line communication system for multiple channel access," in *Proc. 2009 IEEE International Symposium on Power Line Communications and Its Applications*, Dresden, 2009, pp. 47–52.
- [19] J. A. C. Bingham, "Multicarrier modulation for data transmission: an idea whose time has come," *IEEE Communications Magazine*, vol. 28, no. 5, pp. 5–14, May 1990.
- [20] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications," *IEEE Signal Processing Magazine*, vol. 17, no. 3, pp. 29–48, May 2000.
- [21] N. Pavlidou, A. J. Han Vinck, J. Yazdani, and B. Honary, "Power line communications: state of the art and future trends," *IEEE Communications Magazine*, vol. 41, no. 4, pp. 34–40, Apr. 2003.
- [22] A. Dubey, R. K. Mallik, and R. Schober, "Performance analysis of a multi-hop power line communication system over log-normal fading in presence of impulsive noise," *IET Commun.*, vol. 9, no. 1, pp. 1–9, Jan. 2015.
- [23] A. Dubey and R. K. Mallik, "PLC system performance with AF relaying," *IEEE Trans. Commun.*, vol. 63, no. 6, pp. 2337–2345, Jun. 2015.
- [24] A. Dubey and R. K. Mallik, "Effect of channel correlation on multi-hop data transmission over power lines with decode-and-forward relays," *IET Communications*, vol. 10, no. 13, pp. 1623–1630, Jan. 2016.
- [25] K. L. Heo, S. M. Cho, J. W. Lee, M. H. Sunwoo, and Seong Keun Oh, "Design of a high speed OFDM modem system for powerline communications," in *Proc. IEEE Workshop on Signal Processing Systems*, San Diego, CA, USA, 2002, pp. 264–269.
- [26] W. Zhu, X. Zhu, E. Lim and Y. Haung, "State-of-art power line communication's channel modelling," *Procedia Computer Science*, vol. 17, pp. 563–570, 2013.
- [27] P. P. Vaidyanathan, "Multirate Systems and Filter Banks", E. Cliffs, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.
- [28] Daubechies, "Ten lectures on wavelets," *CBMS-NSF Regional Conference Series in Applied Mathematics*, vol. 61, 1992.
- [29] A. Cohen, I. Daubechies, and J.-C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, Vol. 45, No. 5, pp. 485–560, Jun. 1992.
- [30] J. J. van de Beek, M. Sandell, and P. O. Borjesson, "ML estimation of time and frequency offset in OFDM systems," *IEEE Transactions on Signal Processing*, vol. 45, no. 7, pp. 1800–1805, Jul. 1997.

Codebook based Precoding for Multiuser MIMO Broadcast Systems: An MM Approach

Sai Subramanyam Thoota
Dept. of ECE
IISc Bangalore, India
Email: thoota@iisc.ac.in

Prabhu Babu
Centre for Applied Research
in Electronics, IIT Delhi, India
Email: prabhbabu@care.iitd.ac.in

Chandra R. Murthy
Dept. of ECE
IISc Bangalore, India
Email: cmurthy@iisc.ac.in

Abstract—The goal of this paper is to propose a novel, principled approach to solve non-convex optimization problems that arise in multiuser (MU) multiple input multiple output (MIMO) cellular wireless communication systems. We explore a minorization-maximization (MM) optimization approach, which is guaranteed to converge to a stationary point starting from any initialization. One of the important problems in wireless communications is sum rate maximization in MU MIMO broadcast systems, in which multiple data streams are simultaneously transmitted to all users. In this paper, we adopt a codebook based precoding method, where each data stream is beamformed using a vector selected from a predetermined codebook. Our objective is to determine the selection of beamforming vectors and power allocation to each beam to maximize the achievable sum rate. We reformulate the problem to facilitate the application of MM procedure in a nested fashion. The outcome is a novel, iterative, and computationally efficient solution, which we call the inverse-MM (IMM) algorithm. We illustrate the superior performance of our algorithm compared to existing approaches through Monte Carlo simulations. The advantages of computational efficiency, simple implementation, and structured approach makes the MM framework a good candidate for solving non convex optimization problems in wireless communications.

Index Terms—Minorization-Maximization, MU-MIMO, Beamforming, Precoding, Frequency-Division Duplex.

I. INTRODUCTION

Sum rate maximization in cellular downlink (DL) multiuser (MU) multiple input multiple output (MIMO) systems is a well researched area [1]–[5]. In particular, performance optimization with data precoding using a codebook of beamforming vectors has been considered under perfect and imperfect channel state information (CSI) at the base station (BS) [6]–[8]. Codebook based precoding is part of cellular standards like long term evolution (LTE) and LTE-advanced (LTE-A) [9], as it reduces the DL overhead of conveying the precoder matrices in certain transmission modes.¹ Also, in millimeter wave (mmWave) and massive MIMO hybrid precoding [10], RF precoders are implemented in analog hardware, which imposes constraints on the different phase shifts that can be generated. This set of constrained RF phase shifters can be included in a codebook, and the appropriate vectors are chosen by the BS based on the channel conditions.

In this work, we address the problem of joint precoding vector selection and power allocation for sum rate maximization in MU MIMO DL systems. We consider a codebook based digital precoding based on CSI available at the BS. Codebook constrained multi-stream data transmission is intrinsically harder than the (unconstrained) design of precoding matrices, as the underlying problem becomes one of allocating beamforming vectors to users, i.e., an integer optimization problem. Further, due to the inter-stream and inter-user interference, the power allocation problem is non-convex. The problem is thus both combinatorial and non-convex, making it hard to solve. We propose a principled solution to this problem, based on the minorization-maximization (MM) approach.

We assume an explicit channel state feedback² method [11], where quantized CSI is available at the BS. To acquire CSI, the BS first transmits common pilot symbols, using which, the UE estimate their channels, and transmit a quantized version of it back to the BS. The BS chooses beamforming vectors and corresponding powers to maximize the achievable sum rate, and conveys the codebook indices to the UEs. We assume the availability of perfect CSI at the UEs from the initial common pilots sent by the BS, and that the feedback link is delay and error free. This not only simplifies the exposition, but also helps to isolate the performance loss due to the finite-sized codebook at the BS and the finite rate feedback of CSI from the UEs. However, we note that our framework can be extended to the case of estimated CSI at the UEs by appropriately modifying the objective function. We propose an MM-based algorithm to solve the sum-rate optimum beamforming vector selection and power allocation problem. As we will illustrate through numerical simulations, our proposed solution offers excellent performance, is computationally efficient, and exhibits fast convergence.

When using an MM approach for optimizing a non-convex

²Implicit feedback includes the precoding matrix, rank, and channel quality indicators from the UE to the BS. In some scenarios, e.g., low mobility, explicit feedback is more appropriate than implicit feedback: since the coherence interval is large, the BS can acquire CSI and use it over multiple subframes. Also, in frequency division duplex systems, the CSI acquired by the UEs via DL training is quantized and sent back to the BS, which then uses all the UEs' CSI to jointly allot precoding matrices to the UEs from a codebook. Then, the BS only needs to transmit the codebook indices to the UEs over a control channel, to enable data decoding at the receivers.

¹For e.g., in the so-called transmission mode 6 defined in LTE-A [9].

objective function, the key challenge is to bound the cost function by a surrogate cost function that is tight at the current iterate and is easy to optimize (e.g., without involving matrix inverses), thereby leading to a computationally efficient algorithm. Such an algorithm inherits the MM's guaranteed convergence to a local optimum from any initialization. In this paper, we propose an approach satisfying all these properties.

Our solution uses a nested application of MM, in which we lower bound the objective function multiple times to obtain a *computationally efficient, closed form* solution. We choose the lower bounding surrogate functions using new matrix inequalities derived in this paper. These inequalities not only facilitate the use of a nested MM approach to optimization, but are interesting in their own right, as they are potentially useful for solving other non-convex optimization problems.

The main contributions of this paper are:

- 1) Inverse MM (IMM) algorithm: In this procedure, we obtain a surrogate quadratic function using a matrix lemma, which is used to lower bound the non-convex objective function. The surrogate quadratic cost function admits a closed-form optimal solution.
- 2) We compare the performance of the algorithm against the state of the art weighted minimum mean squared error (WMMSE) algorithm [4]. We empirically show that the proposed algorithm outperforms the WMMSE algorithm by a large margin, in terms of the achieved sum rate.

As mentioned earlier, in order to arrive at a tractable optimization problem, we need to judiciously bound the cost function multiple times. This careful choice of the bounding functions is a key novelty of this work.

II. SYSTEM MODEL & PROBLEM STATEMENT

We consider a single-cell frequency division duplexing (FDD) MU MIMO system comprised of a BS equipped with N_t antennas and K users each equipped with N_r antennas. The UEs and the BS share a predetermined codebook $\mathbf{C} \in \mathbb{C}^{N_t \times N}$, whose columns consist of N unit-norm beamforming vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N$, with $\mathbf{v}_j \in \mathbb{C}^{N_t \times 1}$. The complex baseband channel between the BS and the k^{th} UE is denoted by $\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t}$. We assume that the channel is accurately estimated at the UEs using common pilots sent on the DL by the BS. The UEs quantize the CSI using a finite number of bits and send the quantized CSI back to the BS over a control channel. This quantized CSI at the transmitter (CSIT) is used by the BS to allocate beamforming vectors and corresponding data transmit powers to the users. The BS transmits the data symbol $s_k(l)$ to the k^{th} UE by precoding it using the l^{th} beamforming vector \mathbf{v}_l . Thus, the composite data signal $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$ transmitted by the BS to the K UEs is given by

$$\mathbf{x} = \sum_{k=1}^K \sum_{l=1}^N \mathbf{v}_l s_k(l). \quad (1)$$

Note that this general model allows multiple users to receive data on the same beamforming vector or multiple beamforming vectors to be assigned to a given user. Ultimately,

the solution to the sum rate maximization problem will ensure that the allocation maximizes the objective function. Therefore, there is no need to explicitly impose constraints such as each beamforming vector should be allotted to at most one user. In the numerical evaluation, we observe that the same beamforming vector does not get assigned to multiple users. This is intuitive, as the same vector being assigned to multiple users will result in high inter-user interference.

In (1), the data symbols $\{s_k(l)\}$, $k = 1, \dots, K$, $l = 1, \dots, L$ are assumed to be i.i.d. Gaussian distributed, with power $P_k(l) = E(|s_k(l)|^2)$. Let $\Phi_k \triangleq \text{diag}(P_k(1), P_k(2), \dots, P_k(N))$ denote a diagonal matrix whose entries contain the powers allotted to the k^{th} user on the N beamforming vectors. Then, the goal at the BS is to determine Φ_k , based on \mathbf{H}_k , $k = 1, 2, \dots, K$, to maximize the achievable sum rate in the system. Note that, $P_k(l) = 0$ is equivalent to *not allotting* the l^{th} beamforming vector in the codebook to the k^{th} user.

The first step towards solving the above problem is to determine the achievable rate with an arbitrary power allocation matrix Φ_k , $k = 1, 2, \dots, K$. Now, the received signal $\mathbf{y}_k \in \mathbb{C}^{N_r \times 1}$ at the k^{th} user is given by

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{w}_k, \quad (2)$$

where $\mathbf{w}_k \in \mathbb{C}^{N_r \times 1}$ is the complex additive white Gaussian noise of the k^{th} user with distribution $\mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_r})$. The received signal can also be expressed as follows:

$$\mathbf{y}_k = \mathbf{H}_k \sum_{j=1}^K \mathbf{C} \mathbf{s}_j + \mathbf{w}_k, \quad (3)$$

where $\mathbf{s}_j = [s_j(1), s_j(2), \dots, s_j(N)]^T$. The rate achievable for the k^{th} user is given by [3]

$$R_k = \log |\mathbf{I}_{N_r} + \mathbf{V}_k^{-1} \mathbf{H}_k \mathbf{C} \Phi_k \mathbf{C}^H \mathbf{H}_k^H|, \quad (4)$$

where $|\cdot|$ denotes the determinant, and

$$\mathbf{V}_k \triangleq \sigma^2 \mathbf{I}_{N_r} + \sum_{\substack{j=1 \\ j \neq k}}^K \mathbf{H}_k \mathbf{C} \Phi_j \mathbf{C}^H \mathbf{H}_k^H \quad (5)$$

is the interference plus noise covariance matrix.

The DL sum rate is given by $R_{\text{tot}} = \sum_{k=1}^K R_k$. Our goal is to maximize the sum rate R_{tot} under a maximum total power constraint:

$$\begin{aligned} & \underset{\substack{\Phi_1, \Phi_2, \dots, \Phi_K \\ \Phi_k \text{ diagonal, p.s.d.}}}{\text{maximize}} \sum_{k=1}^K \log |\mathbf{I}_{N_r} + \mathbf{V}_k^{-1} \tilde{\mathbf{H}}_k \Phi_k \tilde{\mathbf{H}}_k^H|, \\ & \text{subject to } \text{Tr} \left(\sum_{k=1}^K \Phi_k \right) \leq P_{\text{max}} \end{aligned} \quad (6)$$

where $\tilde{\mathbf{H}}_k = \mathbf{H}_k \mathbf{C}$, and P_{max} is the total transmit power allowed at the BS.

Remark: The rate in (6) is achievable because the desired signal and interference covariance matrices can be computed at the UEs using their CSI \mathbf{H}_k and the power allocation

matrices Φ_k that are made available to them via a DL control channel. This is true even though the BS uses the quantized version of \mathbf{H}_k that is available from the feedback channel in (6) to optimize Φ_k . Note that, for performance evaluation, we use the true \mathbf{H}_k in (6) to compute the sum rate achieved under the prescribed beamforming vector and power allocation.

The optimization problem in (6) is nonconvex, and cannot be solved in closed-form. We also note that it is fundamentally different from the problem of precoder design for sum rate maximization, e.g., the problem considered in [3], [4]. In our case, the matrices Φ_k are restricted to be diagonal and p.s.d., while the past work only requires them to be p.s.d. The additional freedom available in choosing Φ_k makes the latter problem easier to solve. The restriction of the power allocation matrices to be diagonal is required for it to be implementable under codebook based precoding. This constrains the precoding matrices to belong to the set of matrices that can be expressed as the sum of outer products of codebook vectors weighted by the corresponding power allocation, and makes the problem challenging.

In the next section, we propose an iterative algorithm based on the MM principle. An excellent tutorial on MM can be found in [12].

III. PROPOSED ALGORITHM

We now present our solution to the beamforming vector selection and power allocation problem stated in (6). The first step in finding a computationally efficient solution to (6) is to find a surrogate function which is a lower bound on the sum rate, and is tight at the current iterate. The following Lemma presents such a lower bound. The complete proof is omitted due to lack of space.

Lemma 1: For matrices $\mathbf{Z}, \mathbf{Y} \succeq 0$, the non-convex function $f(\mathbf{Z}, \mathbf{Y}) = \log |\mathbf{Z}^{-1}\mathbf{Y}|$ can be lower bounded by

$$f(\mathbf{Z}, \mathbf{Y}) \geq -\left(\log |\mathbf{Z}^{(0)}| + \text{Tr}(\mathbf{Z}^{(0)-1}(\mathbf{Z} - \mathbf{Z}^{(0)}))\right) + \log |\mathbf{Y}^{(0)-1}| + \text{Tr}(\mathbf{Y}^{(0)}(\mathbf{Y}^{-1} - \mathbf{Y}^{(0)-1})) \quad (7)$$

with equality at $\mathbf{Z} = \mathbf{Z}^{(0)}$ and $\mathbf{Y} = \mathbf{Y}^{(0)}$, and $\mathbf{Z}^{(0)-1}, \mathbf{Y}^{(0)-1}$ are the inverses of the matrices $\mathbf{Z}^{(0)}, \mathbf{Y}^{(0)}$, respectively.

Proof: The function f is convex in \mathbf{Z} and \mathbf{Y}^{-1} . Hence, we can bound it from below using the first order Taylor series expansion, resulting in the lower bound given by (7). ■

Next, we define an intermediate matrix

$$\mathbf{B}_k \triangleq \sigma^2 \mathbf{I}_{N_r} + \sum_{j=1}^K \tilde{\mathbf{H}}_k \Phi_j \tilde{\mathbf{H}}_k^H. \quad (8)$$

The rate of the k^{th} user can then be written as $R_k = \log |\mathbf{V}_k^{-1}\mathbf{B}_k|$. This is in the same form as in Lemma 1. Hence, we get the following surrogate optimization problem for (6):

$$\begin{aligned} & \{\Phi_1^{(m+1)}, \dots, \Phi_K^{(m+1)}\} = \\ & \arg \max_{\Phi_1, \dots, \Phi_K} \sum_{k=1}^K \left\{ -\text{Tr}(\mathbf{V}_k^{(m)-1} \sigma^2 \mathbf{I}_{N_r} + \sum_{\substack{j=1 \\ j \neq k}}^K \tilde{\mathbf{H}}_k \Phi_j \tilde{\mathbf{H}}_k^H) \right\} \end{aligned}$$

$$\begin{aligned} & -\text{Tr} \left(\mathbf{B}_k^{(m)} [\sigma^2 \mathbf{I}_{N_r} + \sum_{j=1}^K \tilde{\mathbf{H}}_k \Phi_j \tilde{\mathbf{H}}_k^H]^{-1} \right) \}, \\ & \text{subject to } \text{Tr} \left(\sum_{k=1}^K \Phi_k \right) \leq P_{\max}, \end{aligned} \quad (9)$$

where m is the iteration index. Note that we have omitted the constant terms from (7) in writing the above objective function. Also, in (9), the quantities $\mathbf{V}_k^{(m)}$ and $\mathbf{B}_k^{(m)}$, which are the values of \mathbf{V} and \mathbf{B} in the m^{th} iteration, are computed by substituting $\Phi_k^{(m)}$ for Φ_k in (5) and (8), respectively. Now, if we are able to solve the surrogate problem in (9), then, starting from an arbitrary initialization for Φ_k , the MM procedure iterates between solving (9) and updating \mathbf{V}_k and \mathbf{B}_k . By virtue of the fact that the cost function in (6) increases in each iteration and is bounded above (for example, by the sum of the best rates achievable by each individual user), such a procedure is guaranteed to converge to a local optimum from any initialization.

Now, the optimization problem in (9) is a semidefinite program. However, the matrices to be optimized, $\{\Phi_k\}_{k=1}^K$ are coupled in the objective function and constraints, making it a large dimensional problem. Hence, there is a need to find alternative, computationally inexpensive approaches to solving (9). The proposed inverse MM (IMM) algorithm employs a surrogate function to further lower bound the objective function, leading to a surrogate cost function that is more amenable to optimization. In fact, we are able to solve the final surrogate problem in closed-form. Let

$$\Phi \triangleq \text{diag}(\Phi_1, \dots, \Phi_K) \quad (10)$$

denote the augmented power allocation matrix, and

$$\mathbf{Q} \triangleq \sum_{k=1}^K \text{diag} \left(\tilde{\mathbf{H}}_k^H \mathbf{V}_k^{-1} \tilde{\mathbf{H}}_k, \dots, \mathbf{0}_N, \dots, \tilde{\mathbf{H}}_k^H \mathbf{V}_k^{-1} \tilde{\mathbf{H}}_k \right). \quad (11)$$

In the above, the $N \times N$ all zero matrix $\mathbf{0}_N$ is in the k^{th} block diagonal position of \mathbf{Q} . After omitting the constant terms, the first term in (9) can be rewritten as $\text{Tr}(\mathbf{Q}^{(m)} \Phi)$, where the superscript m denotes the iteration index, and $\mathbf{Q}^{(m)}$ is obtained by substituting $\mathbf{V}_k^{(m)}$ for \mathbf{V}_k in (11).

We define an augmented covariance matrix $\tilde{\Phi} \in \mathbb{R}^{(KN+N_r) \times (KN+N_r)}$, an augmented channel matrix $\tilde{\Psi}_k \in \mathbb{C}^{N_r \times (KN+N_r)}$ and the matrix $\Xi_k \in \mathbb{C}^{N_r \times N_r}$ as follows:

$$\tilde{\Phi} \triangleq \text{diag}(\Phi_1, \dots, \Phi_K, \sigma^2 \mathbf{I}_{N_r}), \quad (12)$$

$$\tilde{\Psi}_k \triangleq [\tilde{\mathbf{H}}_k, \dots, \tilde{\mathbf{H}}_k, \mathbf{I}_{N_r}], k = 1, \dots, K, \quad (13)$$

$$\Xi_k \triangleq \tilde{\Psi}_k \tilde{\Phi} \tilde{\Psi}_k^H, k = 1, \dots, K. \quad (14)$$

In (13), the matrix $\tilde{\mathbf{H}}_k$ is repeated K times. Using the above notation, we can rewrite the term inside the square brackets in (9) as $\mathbf{B}_k^{(m)} \Xi_k^{-1}$.

In order to develop the IMM procedure, we recall the following proposition from [13].

Proposition 1: Let \mathbf{R} be an $m \times n$ matrix and \mathbf{A} be

an $m \times m$ p.s.d. matrix. Then, the function $f(\mathbf{U}) \triangleq \text{Tr} \left(\mathbf{A} \left(\mathbf{R} \mathbf{U} \mathbf{R}^H \right)^{-1} \right)$ can be upper bounded as

$$f(\mathbf{U}) \leq \text{Tr} \left(\mathbf{A} \left(\mathbf{R} \mathbf{U}^{(m)} \mathbf{R}^H \right)^{-1} \mathbf{R} \mathbf{U}^{(m)} \mathbf{U}^{-1} \mathbf{U}^{(m)} \mathbf{R}^H \left(\mathbf{R} \mathbf{U}^{(m)} \mathbf{R}^H \right)^{-1} \right), \quad (15)$$

with equality at $\mathbf{U} = \mathbf{U}^{(m)}$.

Now, since $\mathbf{B}_k^{(m)} \succeq 0 \forall k$, we can apply proposition 1 to $\text{Tr} \left(\mathbf{B}_k^{(m)} \boldsymbol{\Xi}_k^{-1} \right)$, which results in the bound:

$$\begin{aligned} & \sum_{k=1}^K \text{Tr} \left(\mathbf{B}_k^{(m)} \boldsymbol{\Xi}_k^{-1} \right) \\ & \leq \sum_{k=1}^K \text{Tr} \left(\mathbf{B}_k^{(m)} \boldsymbol{\Xi}_k^{(m)-1} \tilde{\Psi}_k \tilde{\Phi}^{(m)} \tilde{\Phi}^{-1} \tilde{\Phi}^{(m)} \tilde{\Psi}_k^H \boldsymbol{\Xi}_k^{(m)-1} \right) \\ & = \text{Tr} \left(\sum_{k=1}^K \tilde{\Phi}^{(m)} \tilde{\Psi}_k^H \boldsymbol{\Xi}_k^{(m)-1} \tilde{\Psi}_k \tilde{\Phi}^{(m)} \tilde{\Phi}^{-1} \right), \end{aligned} \quad (16)$$

where (16) is obtained by recognizing that $\mathbf{B}_k^{(m)} \boldsymbol{\Xi}_k^{(m)-1}$ is the identity matrix, cyclically permuting the terms, and pulling the summation over k into the trace function. Let

$$\mathbf{Z} \triangleq \sum_{k=1}^K \tilde{\Phi} \tilde{\Psi}_k^H \boldsymbol{\Xi}_k^{-1} \tilde{\Psi}_k \tilde{\Phi}. \quad (17)$$

Substituting $\text{Tr} (\mathbf{Q}^{(m)} \Phi)$ (below (11)) and (16) into (6), we get the following surrogate optimization problem:

$$\Phi^{(m+1)} = \arg \min_{\Phi \succeq 0} \left\{ \text{Tr} \left(\mathbf{Q}^{(m)} \Phi + \mathbf{Z}^{(m)} \tilde{\Phi}^{-1} \right) \right\} \quad (18)$$

subject to $\text{Tr} (\Phi) \leq P_{\max}$,

where m is the iteration index.

The above cost function is essentially quadratic in Φ , and is therefore amenable to optimization. Further, the problem in (18) is separable in terms of the optimization variables, and can be solved using the Lagrangian method to obtain a solution in closed form. The Lagrangian is given by

$$\begin{aligned} & \sum_{i=1}^{KN} \left(\left[\mathbf{Q}^{(m)} \right]_{(i,i)} P(i) + \left[\mathbf{Z}^{(m)} \right]_{(i,i)} \frac{1}{P(i)} \right) \\ & + \eta \left(\sum_{i=1}^{KN} P(i) - P_{\max} \right). \end{aligned} \quad (19)$$

From the above, it is easy to show that the solution to the surrogate problem is given by

$$P(i) = \left(\frac{\left[\mathbf{Z}^{(m)} \right]_{(i,i)}}{\left[\mathbf{Q}^{(m)} \right]_{(i,i)} + \eta} \right)^{\frac{1}{2}}, \quad (20)$$

$i = 1, 2, \dots, KN$. Here, η is chosen to satisfy the constraint $\sum_{i=1}^{KN} P(i) = P_{\max}$. Since (20) is strictly decreasing in η , we can determine it using a simple line search.

Thus, in each iteration, we bound the cost function using

Algorithm 1 IMM

Input: $\mathbf{H}_1, \dots, \mathbf{H}_K, \mathbf{C}, K, P_{\max}, \sigma$

Output: $P_1(1), \dots, P_1(N), \dots, P_K(1), \dots, P_K(N)$

- 1: Initialize $P_1(1), \dots, P_1(N), \dots, P_K(1), \dots, P_K(N)$ to satisfy the total power constraint.
 - 2: Compute $\tilde{\mathbf{H}}_k = \mathbf{H}_k \mathbf{C}$, $k = 1, 2, \dots, K$.
 - 3: Compute $\tilde{\Psi}_1, \dots, \tilde{\Psi}_K$ using (13).
 - 4: **repeat**
 - 5: Compute $\Phi, \tilde{\Phi}$ using (10), (12), respectively.
 - 6: Compute \mathbf{Q} and \mathbf{Z} using (11) and (17) respectively.
 - 7: Calculate Lagrange multiplier η using line search to satisfy maximum power constraint P_{\max} .
 - 8: Compute $P(i)$ using (20), $i = 1, 2, \dots, KN$.
 - 9: **for** $k = 1$ to K **do**
 - 10: **for** $i = 1$ to N **do**
 - 11: Compute $P_k(i)$ using (21).
 - 12: **end for**
 - 13: **end for**
 - 14: **until** convergence
-

TABLE I.
STORAGE AND FLOP COUNT ORDER OF IMM PER ITERATION

Matrix	Size	Flop Count
$\boldsymbol{\Xi}_k$	$N_r \times N_r$	$KN N_r^2$
\mathbf{Z}	$(KN + N_r)$ $\times (KN + N_r)$	$KN N_r^2$
\mathbf{Q}	$KN \times KN$	$(N_r + K)N^2$

Proposition 1, and maximize the corresponding surrogate cost function using the solution in (20). Then, we recompute the bounding function, and the process repeats till the cost function converges. The pseudo-code for the IMM procedure is given in Algorithm 1.

The outcome of the IMM procedure is the matrix Φ , which gives the individual users' powers across all the beamforming vectors. The power allotted to the k^{th} user on the j^{th} beamforming vector can be written using the solution from the IMM procedure as

$$P_k(j) = P((k-1)N + j). \quad (21)$$

A. Computational Complexity

We use floating point operations (flops) as the metric to quantify the computational complexity of the proposed algorithm. The per-iteration computational complexity of the IMM algorithm is provided in Table I. The flop count values are calculated after careful analysis of the structural properties of the various matrices. The overall computational complexity of each iteration of the IMM algorithm is $\mathcal{O}((K + N_r)N^2)$, and the storage complexity is $(KN + N_r)^2$.

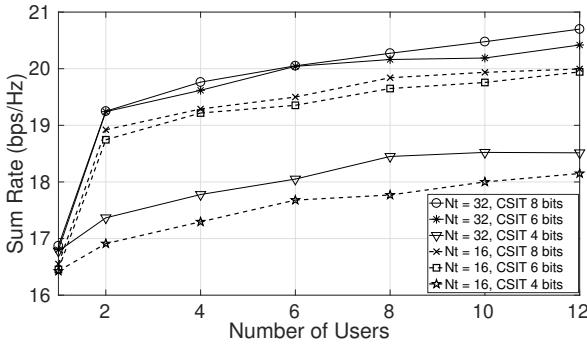


Figure 1. IMIM: Sum rate vs. K , codebook size = 512, for different number of transmit antennas and number of quantization bits for obtaining CSI at the BS.

IV. SIMULATION RESULTS

In this section, we numerically evaluate the performance of the IMIM algorithm. The beamforming codebook used for the initial set of simulations is uniformly distributed on the N dimensional complex unit sphere. The channel coefficients are drawn i.i.d. from $\mathcal{CN}(0, 1)$. The circularly symmetric complex Gaussian additive noise at the receivers is distributed as $\mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_r})$. The algorithm is initialized randomly, and run till convergence. We consider the codebook size varying from $N = 64$ (6 bits) to 1024 (10 bits). We consider $N_t = 16, 32$, while the number of receive antennas is set as $N_r = 2$. The number of users is varied from 1 to 12. For the feedback of CSI from the UEs to the BS, we quantize the CSI to 4, 6, and 8 bits per matrix entry. The noise variance is set to unity, and the maximum transmit power is 20 dB above the noise floor.

We compare the sum rate of the IMIM algorithm against the state of the art WMMSE algorithm [4]. This algorithm assumes perfect CSIT for designing the precoder matrix, but does not use codebook based beamforming to maximize the sum rate. Hence, for fair comparison with our work, we quantize the precoding vectors output by the above algorithm to the corresponding vectors in the codebook with the largest magnitude inner product, and compare the sum rate achieved. This ensures that the DL control overhead is the same as that of our codebook-based precoding approach.

Figure 1 shows the sum rate of IMIM vs. the number of users with a codebook size of 512 (9 bits). As the number of users increases, the multiuser interference increases, but the overall sum rate continues to increase as long as the number of allotted spatial streams is smaller than the total number of spatial dimensions available. Hence, it saturates beyond about 8 users for $N_t = 16$. The sum rate also increases with the number of quantization bits of the CSIT, but the sum rate increase from 6 to 8 quantization bits of CSIT is minimal. This performance plot also serves as a benchmark to determine the number of quantization bits needed to achieve a certain quality of service.

Figure 2 shows the sum rate performance of the state of the art WMMSE algorithm. We have plotted the sum rates obtained by directly applying the WMMSE solution as well

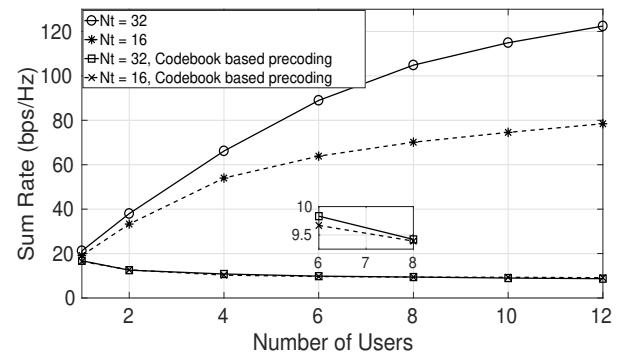


Figure 2. WMMSE: Sum rate vs. K , codebook size = 1024, with the CSIT quantized using 8 bits. Sum rate achieved by the IMIM algorithm with a codebook size of 512 outperforms the sum rate achieved by WMMSE with a codebook size of 1024 as seen in Fig. 1.

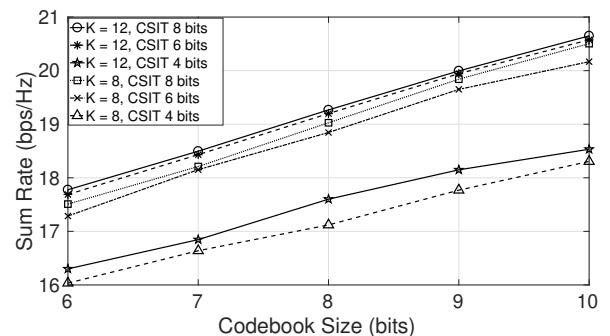


Figure 3. IMIM: Sum rate vs. codebook size (in bits). The sum rate increases linearly with the codebook size.

as the sum rates obtained by quantizing the precoding vectors output by the WMMSE algorithm to the vector in codebook with maximum magnitude inner product with each precoding vector. We see that the sum rate decreases dramatically when the WMMSE algorithm is directly adapted to codebook-based precoding, resulting in an over two-fold lower rate than the proposed IMIM algorithm (See also Fig. 3). Moreover, $N_t = 16$ and $N_t = 32$ offer almost the same sum rate, and this sum rate decreases marginally as the number of users increases. Not only is the WMMSE not able to exploit the additional transmit degrees of freedom available, the quantization of beamforming vectors without considering the resulting higher inter-user interference is highly suboptimal. This shows that algorithms designed to maximize the sum rate without considering the codebook constraints are not suitable for a codebook based precoding scenario.

In Fig. 3, we plot the sum rate vs. the codebook size in bits for $N_t = 16$ and the number of users $K = 8, 12$. The sum rate improves roughly linearly with the codebook size, as the transmitter has more choices for assigning beamforming vectors to users. As observed earlier, the sum rate increase from 6 bits quantization of CSIT to 8 bits is negligible. Also, the sum rate increases only marginally as the number of users increases. This is because, with $K = 8$ users with $N_r = 2$ antennas each and $N_t = 16$ antennas at the BS, all the

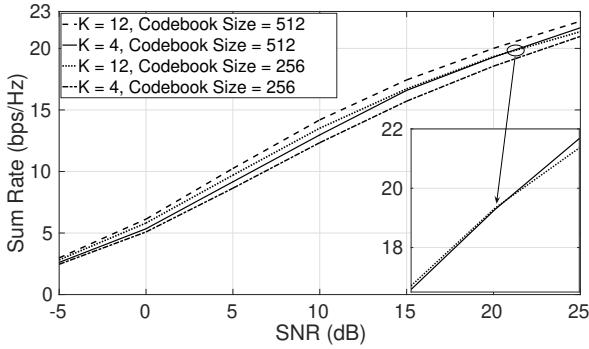


Figure 4. Sum Rate vs. SNR (dB). The sum rate performance of $K = 4, N = 512$ is better than the $K = 12, N = 256$ for SNR greater than 20 dB.

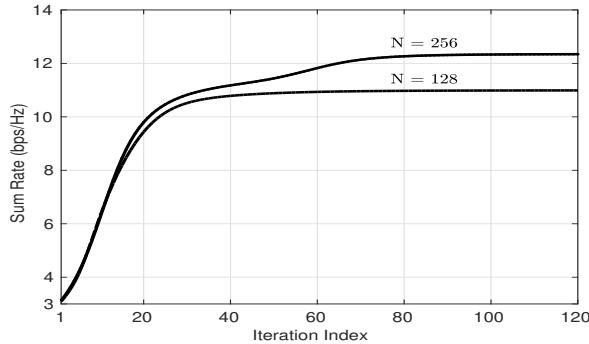


Figure 5. IMM convergence, $K = 4$, $\text{SNR} = 10$ dB. The monotonic convergence behavior is clear from the plot, and the number of iterations to converge increases roughly linearly with codebook size.

spatial degrees of freedom are used up in transmitting to the K UEs. Hence, the improvement in rate from increasing $K = 8$ to $K = 12$ is only marginal. Ideally, i.e., with a precoder that is not constrained by a codebook, the sum rate should saturate beyond 8 users. However, due to the codebook based precoding, we observe a marginal increase in the sum rate even beyond 8 users.

Figure 4 shows the sum rate performance of the IMM algorithm as a function of the SNR, with 4 and 12 users and codebook sizes of 256 and 512. The receiver uses 8 bit quantization of the CSI for feedback to the transmitter. While the sum rate increases with SNR, the performance of all four cases is nearly the same at $\text{SNR} = -5$ dB. In the low SNR, noise dominated regime, more users can be assigned resources without a significant increase in the multiuser interference. Hence, the sum rate achieved with $K = 12$ is higher than that achieved with $K = 4$ for both codebook sizes. At high SNRs, e.g., beyond 20 dB, the sum rate achieved when $K = 4$ users and codebook size of 512 is higher than that of $K = 12$ users and codebook size of 256. This is because, in the interference-limited scenario, having more choices of beamforming vectors can reduce the multi-user interference, leading to a marginally better achievable rate compared to the $K = 12$ user case.

Figure 5 shows the monotonic convergence of the cost function, a property of the MM procedure. Convergence

occurs within about 40 iterations when $N = 128$, with the number of iterations required slightly increasing as N is increased. This is reasonable, as the complexity of the problem increases dramatically with codebook size.

V. CONCLUSIONS

We proposed an algorithm, named inverse MM (IMM), to solve the codebook based DL sum rate maximization problem in an FDD MU MIMO DL broadcast system. This procedure, which is based on a nested application of the MM principle to lower bound the objective function, finds a locally optimal allocation of beamforming vectors and data signal powers to each user, so as to maximize the DL sum rate. The novelty of the algorithm lies in the choice of the surrogate functions used to bound the objective function. This nested framework can also be potentially used in a variety of other resource allocation problems.

We compared the DL sum rate performance of the proposed algorithm with the state-of-the-art WMMSE method, and found that IMM outperforms the quantized version of the other algorithm by a large margin.

REFERENCES

- [1] J. Park and B. Clerckx, "Multi-user linear precoding for multi-polarized massive MIMO system under imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2532–2547, May 2015.
- [2] E. G. Larsson and H. V. Poor, "Joint beamforming and broadcasting in massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 3058–3070, Apr. 2016.
- [3] S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
- [4] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sep. 2011.
- [5] J. Kaleva, A. Tölli, and M. Juntti, "Decentralized sum rate maximization with QoS constraints for interfering broadcast channel via successive convex approximation," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2788–2802, Jun. 2016.
- [6] S. He, J. Wang, Y. Huang, B. Ottersten, and W. Hong, "Codebook-based hybrid precoding for millimeter wave multiuser systems," *IEEE Trans. Signal Process.*, vol. 65, no. 20, pp. 5289–5304, Oct. 2017.
- [7] Z. Xiao, P. Xia, and X.-G. Xia, "Codebook design for millimeter-wave channel estimation with hybrid precoding structure," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 141–153, Jan. 2017.
- [8] A. Liu and V. K. Lau, "Impact of CSI knowledge on the codebook-based hybrid beamforming in massive MIMO," *IEEE Trans. Signal Process.*, vol. 64, no. 24, pp. 6545–6556, Dec. 2016.
- [9] 3GPP, "Technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); physical layer procedures (release 9)," *Tech. Rep. 36.213 (v9.3.0)*, 2010.
- [10] A. Alkhateeb, G. Leus, and R. W. Heath, "Limited feedback hybrid precoding for multi-user millimeter wave systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6481–6494, Nov. 2015.
- [11] C. Lim, T. Yoo, B. Clerckx, B. Lee, and B. Shim, "Recent trend of multiuser MIMO in LTE-advanced," *IEEE Communications Magazine*, vol. 51, no. 3, pp. 127–135, Mar. 2013.
- [12] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, no. 1, pp. 30–37, 2004.
- [13] Y. Sun, P. Babu and D. P. Palomar, "Robust estimation of structured covariance matrix for heavy-tailed elliptical distributions," *IEEE Trans. Signal Process.*, vol. 64, no. 14, pp. 3576–3590, Jul. 2016.

Modified Generalised Quadrature Spatial Modulation

Kiran Gunde and K.V.S. Hari

Department of ECE, Indian Institute of Science, Bangalore-560012, India

Email: gunde@iisc.ac.in, hari@iisc.ac.in

Abstract—In this paper, we propose a Modified Generalised Quadrature Spatial Modulation (mGQSM) scheme with multiple RF chains. The proposed scheme, compared to GQSM, proposes a novel codebook design which provides one extra bit per channel use (bpcu) spectral efficiency with the constraint of $\{\log_2 \left(\frac{N_t}{N_{rf}}\right)\} \geq 0.5$, where N_t denotes number of transmit antennas, and N_{rf} denotes number of RF chains, $1 \leq N_{rf} \leq \lfloor \frac{N_t}{2} \rfloor$. Using the ML detection algorithm, we study the performance of mGQSM with and without imperfect channel state information, via numerical simulations. We compute the computational complexity of ML-decoding in terms of real valued multiplications and introduce a variant of mGQSM called Reduced Codebook mGQSM (RC-mGQSM) to reduce the complexity but resulting in a decrease in spectral efficiency.

I. INTRODUCTION

Multiple-Input Multiple-Output (MIMO) wireless communication technology like the well-known V-BLAST (Vertical-Bell Laboratories Layered Space-Time) [1], also known as Spatial Multiplexing (SMX) is used for achieving higher data rates. However, the number of Radio Frequency (RF) chains leads to increase in the hardware complexity, power consumption and cost, of the system. Spatial Modulation (SM) [2], [3] is an energy efficient scheme for MIMO wireless communication system where it uses only one RF chain that reduces the hardware complexity and cost of the system. In SM, the information bits are divided into two blocks, one for selecting a modulation symbol from a constellation corresponding to M-QAM and the other block for selecting the antenna index.

The spectral efficiency of SM is $\eta_{SM} = \log_2(N_t M)$ bits per channel use (bpcu), where M denotes order of modulation constellation (M-QAM) and N_t denotes the number of transmit antennas. In SM, $\log_2(M)$ bits select the modulation symbol from the constellation and other $\log_2(N_t)$ bits select only one antenna out of all possible transmit antennas, which avoids inter channel interference (ICI) and inter symbol interference (ISI). The individual error performances of SM symbol, antenna index and transmitted symbol were given in [4]. A simplified SM scheme [5], known as space shift keying (SSK) uses only the index of the active antenna to transmit information. The spectral efficiency of SSK is $\log_2(N_t)$ which is lower compared to that of SM. A variant of SM scheme called Quadrature Spatial Modulation (QSM) [6] use both in-phase and quadrature dimensions to transmit the data symbol in one time instant, using multiple antennas. QSM enhances

the spectral efficiency of SM and is free from ICI due to orthogonality of I and Q channel signals.

Generalised Spatial Modulation (GSM) [7] and Multiple Active Spatial Modulation (MASM) [8] schemes were developed to increase the spectral efficiency of SM by increasing the number of RF chains. In GSM, the same data symbol transmits on different antennas hence, it avoids ISI. In MASM, multiple transmitting antennas transmit different data symbols at one time instant. Recently, Generalised Quadrature Spatial Modulation (GQSM) [9] scheme with antenna grouping was developed to increase the spectral efficiency of QSM by grouping transmit antennas according to QSM principle. This scheme presents the uncoded GQSM error performances with different antenna configurations. In this paper, we propose a modified GQSM (mGQSM) scheme by proposing a novel codebook design resulting in the improvement in spectral efficiency of the scheme.

The rest of this paper is organized as follows. Section II describes the system models, section III, presents the proposed mGQSM system model and the implementation and section IV, presents the simulation results followed by section V presenting the conclusion.

II. SYSTEM MODEL

Consider $N_r \times N_t$ MIMO system model with N_t transmit antennas and N_r receive antennas, and frequency-flat Rayleigh fading channel. Let \mathbf{H} be a complex channel matrix with $N_r \times N_t$ dimension and \mathbf{n} is noise vector with $N_r \times 1$ dimension. The elements of the \mathbf{H} and \mathbf{n} are assumed to be independent and identically distributed (i.i.d.) complex Gaussian distribution $\mathcal{CN}(0, 1)$ and $\mathcal{CN}(0, N_0)$, respectively, where N_0 is noise variance. The received signal at the output of the channel is given by,

$$\mathbf{y} = \sqrt{E_s} \mathbf{H} \mathbf{x} + \mathbf{n} \quad (1)$$

Where E_s denotes transmitted symbol energy and the transmitted vector \mathbf{x} is selected from the codebook \mathcal{C} , i.e., $\mathbf{x} \in \mathcal{C}$ which consist of all possible transmitted vectors .

In SM, the transmitted vector \mathbf{x} can be represented by,

$$\mathbf{x}_{l,s} = [0, \dots, 0, s, 0, \dots, 0]^T \in \mathbb{C}^{N_t} \text{ and } \|\mathbf{x}_{l,s}\|_0 = N_{rf}$$

⁰Notation: Bold lowercase and uppercase letters denote column vectors and matrices, respectively. $\{\cdot\}$ denotes the fractional part of the decimal value, $\|\cdot\|_0$ denotes zero norm of a vector, $|\cdot|$ denotes cardinality of a set, $\lfloor x \rfloor$ denotes integer part of x . $\Re\{\cdot\}$ and $\Im\{\cdot\}$ are denotes real and imaginary coefficients of the complex symbol respectively, (\cdot) denotes binomial coefficient.

where $l \in L = \{i\}_{i=1}^{N_t}$ (set of transmit antenna indices and number of), $s \in S$ (signal set for M-QAM modulation scheme), and N_{rf} denotes as number of RF-chains, $N_{rf} = 1$. The codebook of SM represented as \mathcal{C}_{SM} , which consists of all possible transmitted vectors and $|\mathcal{C}_{SM}| = 2^{\eta_{SM}}$.

In QSM, the transmitted symbol ($s = s_r + js_i$) is divided into real (s_r) and imaginary (s_i) symbols and transmitted over same or different antennas (l_r, l_i), where l_r and l_i are the transmit antenna indices corresponding to s_r and s_i respectively. The transmitted vector \mathbf{x} can be represented in two instants with $N_{rf} = 1$,

- 1) If only one antenna is active, the transmitted vector, $\mathbf{x}_{l_r, l_i, s_r, s_i} = [0, \dots, 0, s_r + js_i, 0, \dots, 0]^T$ and $\|\mathbf{x}_{l_r, l_i, s_r, s_i}\|_0 = N_{rf}$ and $|\mathcal{C}_{QSM1}| = N_t$.
- 2) If two antennas are active then the transmitted vector, $\mathbf{x}_{l_r, l_i, s_r, s_i} = [0, \dots, 0, s_r, 0, \dots, 0, js_i, \dots, 0]^T$ and $\|\mathbf{x}_{l_r, l_i, s_r, s_i}\|_0 = 2N_{rf}$ and $|\mathcal{C}_{QSM2}| = 2^{\eta_{QSM}} - N_t$.

where $\{l_r, l_i\} \in L = \{i\}_{i=1, r=1}^{N_t}$, $s = s_r + js_i \in S$ (signal set for M-QAM modulation scheme) and η_{QSM} denotes the spectral efficiency of QSM. The codebook of QSM is the union of two codebooks, $|\mathcal{C}_{QSM}| = |\mathcal{C}_{QSM1}| + |\mathcal{C}_{QSM2}| = 2^{\eta_{QSM}}$.

The codebook structure of GSM with multiple data symbols is given in [10], the set of all possible transmit vectors is given by, $\mathcal{C}_{MA-SM} = \{\mathbf{x} | s_j \in \mathbb{A}_0, \|\mathbf{x}\|_0 = N_{rf}, \mathcal{I}(\mathbf{x}) \in \mathbb{S}\}$, where $\mathbb{A}_0 \triangleq \mathbb{A}_M \cup \{0\}$ denotes an effective alphabet and \mathbb{A}_M denotes set of M-QAM symbols and $j = 1, 2, \dots, N_t$. \mathbb{S} denotes the antenna activation pattern such that $\mathbf{p} \in \mathbb{S}$ and $\|\mathbf{p}\|_0 = N_{rf}$, $\mathbf{p}_j \in \{0, 1\}$ and \mathbb{S} denotes the set of possible antenna activation patterns , and $\mathcal{I}(\mathbf{x})$ is a function that gives antenna activation pattern for \mathbf{x} .

In GQSM scheme, the total number of transmit antennas N_t are divided into groups, $n_B = \frac{N_t}{2}$, where n_B denotes number of groups. Each group work with one M-QAM modulation symbol and activate one or two antennas according to QSM principle. Therefore, the spectral efficiency of GQSM is $n_B(\log_2 M + 2)$ bpcu. The codebook structure of GQSM scheme is given by , $\mathcal{C}_{GQSM} = \{\mathbf{x} | x_m = x_{mr} + jx_{mi}, x_{mr} \in \Re\{\mathbb{A}_M\}, x_{mi} \in \Im\{\mathbb{A}_M\}, n_B \leq \|\mathbf{x}\|_0 \leq N_t, \mathcal{I}(\mathbf{x}) \in \mathbb{S}\}$ where, \mathbb{A}_M denotes set of M-QAM symbols and $m = 1, 2, \dots, n_B$. \mathbb{S} denotes the antenna activation pattern such that $\mathbf{p} \in \mathbb{S}$, $\mathbf{p}_j \in \{0, 1\}$ and $|\mathbb{S}| = 2^{n_B}$, and $\mathcal{I}(\mathbf{x})$ is a function that gives antenna activation pattern for \mathbf{x} .

III. PROPOSED MODIFIED GENERALISED QUADRATURE SPATIAL MODULATION (mGQSM) SYSTEM

We propose a Modified Generalised Quadrature Spatial Modulation (mGQSM) scheme with multiple RF chains. The proposed mGQSM system relaxes the antenna grouping at the transmitter and allows multiple antenna activation patterns, to increase the spectral efficiency. The proposed scheme provides extra one bpcu spectral efficiency over GQSM scheme with the constraint of $\{\log_2 \binom{N_t}{N_{rf}}\} \geq 0.5, 1 \leq N_{rf} \leq \lfloor \frac{N_t}{2} \rfloor$. QSM is a special case of mGQSM with $N_{rf} = 1$.

Consider $N_r \times N_t$ system model, Let $h_{l_{mr}}$ and $h_{l_{mi}}$ are l^{th} and l^{th} columns of channel matrix \mathbf{H} , respectively, i.e.

$h_{l_{mr}} = [h_{1,l_{mr}}, \dots, h_{N_r,l_{mr}}]^T$, where $m = 1, 2, \dots, N_{rf}$. The received signal is given by

$$\mathbf{y} = \sqrt{E_s} \sum_{m=1}^{N_{rf}} (\mathbf{h}_{l_{mr}} x_{mr} + j \mathbf{h}_{l_{mi}} x_{mi}) + \mathbf{n} \quad (2)$$

where $l_{mr}, l_{mi} = 1, 2, \dots, N_t$, $m = 1, 2, \dots, N_{rf}$ and E_s denotes transmitted energy.

For a given N_t we can design the codebook \mathcal{C}_{mGQSM} , given by $\mathcal{C}_{mGQSM} = \{\mathbf{x} | x_m = x_{mr} + jx_{mi}, x_{mr} \in \Re\{\mathbb{A}_M\}, x_{mi} \in \Im\{\mathbb{A}_M\}, N_{rf} \leq \|\mathbf{x}\|_0 \leq 2N_{rf}, \mathcal{I}(\mathbf{x}) \in \mathbb{S}\}$, where, \mathbb{A}_M denotes a set of M-QAM symbols and $m = 1, 2, \dots, N_{rf}$, \mathbb{S} denotes the antenna activation pattern such that $\mathbf{p} \in \mathbb{S}$ and $N_{rf} \leq \|\mathbf{p}\|_0 \leq 2N_{rf}$, $\mathbf{p}_j \in \{0, 1\}$ and \mathbb{S} denotes the set of antenna activation pattern, and $\mathcal{I}(\mathbf{x})$ is a function that gives antenna activation pattern for \mathbf{x} .

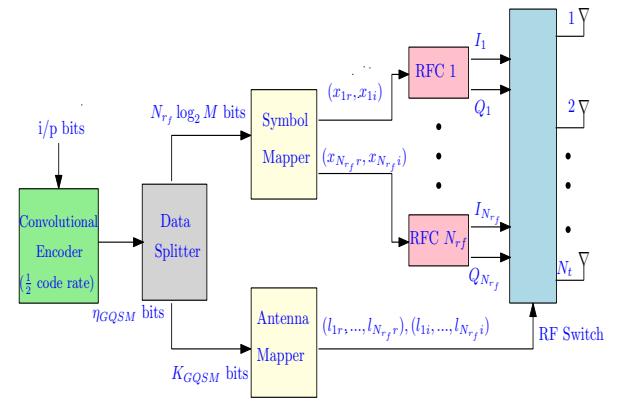


Figure 1. mGQSM Transmitter system diagram.

In mGQSM scheme, multiple data symbols are divided into real and imaginary parts and these parts are transmitted as in-phase and quadrature components by selecting any possible antenna activation patterns available in mGQSM, resulting in the choice of antenna activation patterns being doubled in mGQSM compared to GQSM which yields the extra one bpcu spectral efficiency over GQSM. In mGQSM, the possible antenna activation patterns are equal to $\binom{N_t}{N_{rf}}^2$, we can select any $2^{\lfloor 2 \log_2 \binom{N_t}{N_{rf}} \rfloor}$ antenna activation patterns randomly. Whereas in GQSM scheme, the possible antenna activation patterns are equal to 2^{N_t} and it is half compared to mGQSM. The number of active antennas varies from N_{rf} to $2N_{rf}$ at every symbol period hence, the proposed scheme allows more number of antenna activation patterns to increase the spectral efficiency. The spectral efficiency of mGQSM scheme is given by,

$$\eta_{mGQSM} = \left\lfloor 2 \log_2 \binom{N_t}{N_{rf}} \right\rfloor + N_{rf} \log_2 M \text{ bpcu} \quad (3)$$

Fig. 1 shows the transmitter diagram of the proposed mGQSM scheme with a convolutional encoder. The convolutional encoder with $\frac{1}{2}$ code rate converts the block of $\eta_{C-mGQSM} = \frac{\eta_{mGQSM}}{2}$ input data bits to η_{GQSM} data bits, where $\eta_{C-mGQSM}$ is the spectral efficiency of coded mGQSM. Furthermore, these η_{mGQSM} data bits are divided into two blocks, called the symbol mapper and antenna mapper. The symbol mapper takes $N_{rf} \log_2(M)$ bits and maps them to a complex-valued symbol, from a constellation, and then separates it into In-phase (I) and Quadrature-phase (Q) real-valued symbols. The antenna mapper takes $\left\lfloor 2 \log_2 \left(\frac{N_t}{N_{rf}} \right) \right\rfloor$ bits and maps them to antenna indices which will be selected to transmit the symbols obtained from the symbol mapper block.

We observe that one can select any $2^{K_{mGQSM}}$ antenna active patterns out of $\binom{N_t}{N_{rf}}^2$ patterns as described below, where $K_{mGQSM} = \left\lfloor 2 \log_2 \left(\frac{N_t}{N_{rf}} \right) \right\rfloor$.

A. Antenna Activation Pattern Selection

- 1) Choose N_t .
- 2) Write possible N_{rf} values, using $1 \leq N_{rf} \leq \lfloor \frac{N_t}{2} \rfloor$.
- 3) Select one N_{rf} value.
- 4) Write $\binom{N_t}{N_{rf}}^2$ possible antenna activation pattern vectors and divide into two groups for the transmission of real and imaginary coefficients of Tx symbol.
- 5) Select any antenna activation pattern set \mathbb{S} , which consists of $2^{K_{mGQSM}}$ vectors out of $\binom{N_t}{N_{rf}}^2$ vectors.

For example, consider $N_t = 4$ and $N_{rf} = 2$. The possible number of antenna activation patterns are $\binom{N_t}{N_{rf}}^2 = \binom{4}{2}^2 = 6^2$, $K_{mGQSM} = 5$ and we can select any 2^5 antenna patterns out of 6^2 antenna patterns. It is shown later, via simulations, that the choice of these patterns does not alter the performance of the system. For 4-QAM modulation scheme the transmission rate is 9 bpcu, select [1 0 0 1 0 1 1 0 0] bits for transmission at particular symbol time period. First 5 bits, [1 0 0 1 0] select the antenna activation pattern selection, we have selected antenna active patterns $(l_{1r}, l_{2r}) = (1, 2)$ for real coefficients and $(l_{1i}, l_{2i}) = (2, 3)$ for imaginary coefficients of symbols are to be transmitted. Next $N_{rf} \log_2 M = 4$ bits, [1 1 0 0] select the two 4-QAM modulation symbols. First, [1 1] bits select the symbol $x_1 = +1 - j1$ and remaining, [0 0] bits select the symbol $x_2 = -1 + j1$. Furthermore, these two symbols are divided into real and imaginary coefficients ($x_{1r} = +1, x_{1i} = -1$) and ($x_{2r} = -1, x_{2i} = +1$), respectively. Therefore, the real symbols $x_{1r} = +1$ and $x_{2r} = -1$ are transmitted by antennas $l_{1r} = 1$ and $l_{2r} = 2$, respectively, so the transmitted vector is $\mathbf{x}_r = [+1 -1 0 0]^T$. Similarly, the imaginary symbols $x_{1i} = -1$ and $x_{2i} = +1$ are transmitted by antennas $l_{1i} = 2$ and $l_{2i} = 3$ respectively. So the transmitted vector is $\mathbf{x}_i = [0 -1 +1 0]^T$. The final transmitted vector over the channel is $\mathbf{x} = \mathbf{x}_r + j\mathbf{x}_i = [+1 -1 -j1 +j1 0]^T$.

Receiver Structure:

Maximum-likelihood (ML) detection is used to estimate the antenna indices and symbols transmitted, using

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathcal{C}_{mGQSM}}{\operatorname{argmin}} \| \mathbf{y} - \sqrt{E_s} \mathbf{H} \mathbf{x} \|^2. \quad (4)$$

$$\begin{aligned} & [\hat{l}_{mr}, \hat{l}_{mi}, \hat{x}_{mr}, \hat{x}_{mi}] \\ &= \underset{l_{mr}, l_{mi}, x_{mr}, x_{mi}}{\operatorname{argmin}} \| \mathbf{y} - \sqrt{E_s} \sum_{m=1}^{N_{rf}} (\mathbf{h}_{l_{mr}} x_{mr} + j \mathbf{h}_{l_{mi}} x_{mi}) \|^2 \end{aligned} \quad (5)$$

where \hat{l}_{mr} and \hat{l}_{mi} denote the detected active antenna indices corresponding to \hat{x}_{mr} and \hat{x}_{mi} . \hat{x}_{mr} and \hat{x}_{mi} for $m = 1, 2, \dots, N_{rf}$.

B. Computational complexity and reduced codebook mGQSM

By considering total number of real-valued multiplications, we compute the computational complexity of ML-decoding algorithm for mGQSM. The per-symbol computational complexity, in terms of real-valued multiplications, for the ML solutions is obtained as $10N_r \times 2^{\eta_{mGQSM}}$, where N_r is number of receive antennas and η_{mGQSM} is the spectral efficiency of mGQSM. We can reduce the complexity of a mGQSM scheme by reducing the size of the codebook by considering fewer antenna activation patterns which would also reduce the spectral efficiency, we named this scheme as reduced codebook mGQSM (RC-mGQSM). We derive the spectral efficiency of reduced codebook mGQSM and it is given by,

$$\eta_{RC-mGQSM} = \eta_{mGQSM} - p \text{ bpcu} \quad (6)$$

where η_{mGQSM} is the spectral efficiency of mGQSM and p is an integer value satisfies, $1 \leq p \leq K_{mGQSM}$.

Now, the available antenna pattern vectors, $|\mathbb{S}| = 2^{K_{mGQSM}-p}$ are reduced by increasing the value of p hence, the computational complexity reduces. The computational complexity of SM, QSM, GQSM, mGQSM and RC-mGQSM for $N_t = 4, N_r = 4, N_{rf} = 2, M = 4$ are given in Table I. The spectral efficiency of mGQSM is $\eta_{mGQSM} = 9$ bpcu. If we choose $p = 1$, the number of antenna activation patterns $|\mathbb{S}| = 2^{K_{mGQSM}-p} = 2^4$, which reduces the codebook size and therefore the complexity from 20480 to 10240 real valued multiplications. The computational complexity in GQSM is same as the RC-mGQSM for $p = 1$. Whereas in SM and QSM systems the complexity is very low compared to the mGQSM. However, the spectral efficiency of RC-mGQSM is reduces to $\eta_{mGQSM} - p = 8$ bpcu. The possible Minimum spectral efficiency in RC-mGQSM is 4 bpcu for maximum value of $p = K_{mGQSM} = 5$.

C. mGQSM system with imperfect CSIR

In this section we study the performance of the mGQSM system with imperfect channel state information at the receiver (CSIR) conditions. At the receiver, the ML decoding uses the error channel matrix $\Delta \mathbf{H}$ in addition to the perfect CSIR channel matrix \mathbf{H} for the received signal given in equation 7. The dimension of $\Delta \mathbf{H}$ is same as the dimension of \mathbf{H} that

is $N_r \times N_t$. The entries of \mathbf{H} and $\Delta\mathbf{H}$ are i.i.d $\mathcal{CN}(0, 1)$ and $\mathcal{CN}(0, \sigma_h)$ and respectively, where σ_h is the variance of the imperfect channel. The noise is assumed as i.i.d complex Gaussian random variable with mean zero and variance N_o . For imperfect CSIR conditions, we define signal-to-noise-ratio, SNR as $\frac{E_s(1+\sigma_h)}{N_o}$, where E_s is the transmitted symbol energy and N_o is the noise power. The ML detection of the transmitted signal \mathbf{x} is given by,

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathcal{C}_{mGQSM}}{\operatorname{argmin}} \|\mathbf{y} - \sqrt{E_s}(\mathbf{H} + \Delta\mathbf{H})\mathbf{x}\|^2 \quad (7)$$

where \mathcal{C}_{mGQSM} is denoted as codebook of the mGQSM system and $\hat{\mathbf{x}}$ is denoted as the estimated vector of the transmitted signal vector \mathbf{x} .

Table I
SPECTRAL EFFICIENCY AND COMPUTATIONAL COMPLEXITY
($N_t = 4, N_r = 4, N_{rf} = 2, M = 4$)

Transmission scheme	Spectral efficiency (bpcu)	Computational complexity (real valued multiplications)
SM	$\eta_{SM} = \log_2(N_t M) = 4$	$6N_r \times 2^{\eta_{SM}} = 384$
QSM	$\eta_{QSM} = 2 \log_2 N_t + \log_2 M = 6$	$6N_r \times 2^{\eta_{QSM}} = 1536$
GQSM	$\eta_{GQSM} = \frac{N_t}{2} (\log_2 M + 2) = 8$	$10N_r \times 2^{\eta_{GQSM}} = 10240$
mGQSM	$\eta_{mGQSM} = \left\lceil 2 \log_2 \left(\frac{N_t}{N_{rf}} \right) \right\rceil + N_{rf} \log_2 M = 9$	$10N_r \times 2^{\eta_{mGQSM}} = 20480$
RC-mGQSM	$\eta_{RC-mGQSM} = \eta_{mGQSM} - p = 8(p=1)$	$10N_r \times 2^{\eta_{RC-mGQSM}} = 10240$

IV. SIMULATION RESULTS

In this section, we study the performance of the proposed mGQSM scheme along with QSM, MA-SM and GQSM. We consider frequency-flat Rayleigh fading channel and the channel path gains are assumed to be i.i.d $\mathcal{CN}(0, 1)$. The noise is assumed as i.i.d $\mathcal{CN}(0, N_o)$. We define signal-to-noise-ratio (SNR) as $\frac{E_s}{N_o}$, where E_s is the transmitted symbol energy and N_o is noise power. For all the simulations we use two $\frac{1}{2}$ code rate convolutional encoders, one encoder place at the transmitter and other encoder place after Viterbi decoder at the receiver. We compute the bit-error-ratio (BER) by comparing the output bits of first convolutional encoder at the transmitter and second convolutional encoder at the receiver. We use 10^6 symbols to compute the BER for different SNR values.

In all the figures the triplet denotes (N_t, N_r, N_{rf})

Fig. 2 shows the BER performance comparison of mGQSM and QSM with $N_t = 4$ and $N_r = 4$ system model and the same spectral efficiency with $\eta_{mGQSM} = \eta_{QSM} = 9$ bpcu. We compare the BER performance of mGQSM and QSM with $N_{rf} = 2$ and $N_{rf} = 1$, respectively. For the probability of error 10^{-3} , we observe that the BER performance of mGQSM system with 4-QAM modulation scheme gain 2.5 dB value in SNR over QSM system with 32-QAM modulation scheme.

Fig. 3 shows the BER performance comparison of mGQSM and MA-SM with different system models and the same spectral efficiency with $\eta_{mGQSM} = \eta_{MA-SM} = 9$ bpcu. Let $N_t = 4$ and $N_r = 4$ system model for mGQSM and $N_t = 5$

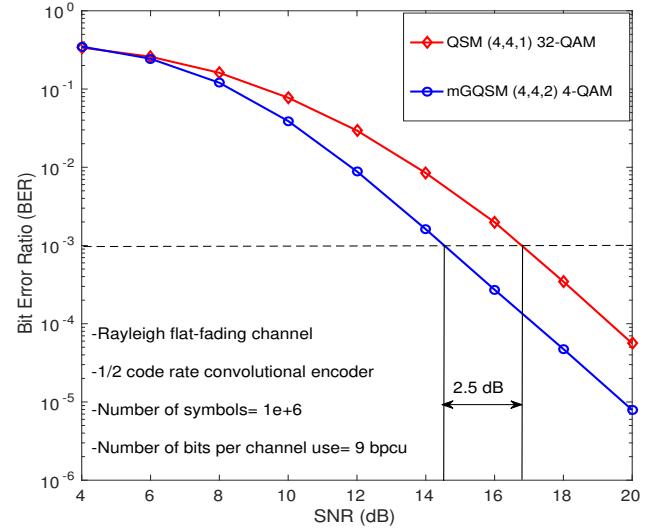


Figure 2. BER performance vs SNR for mGQSM and QSM with $\frac{1}{2}$ code rate convolutional encoder with $N_t = 4$ and $N_r = 4$ system model and the same spectral efficiency with $\eta_{mGQSM} = \eta_{QSM} = 9$ bpcu.

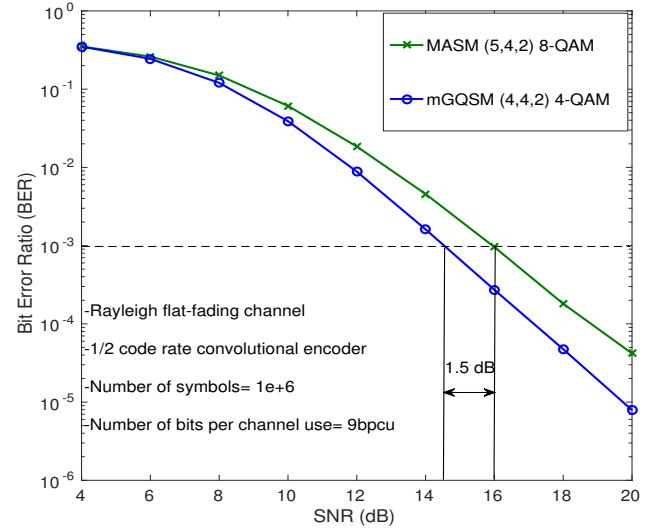


Figure 3. BER performance vs SNR for mGQSM ($N_t = 4, N_r = 4$) and MA-SM ($N_t = 5, N_r = 4$) with $\frac{1}{2}$ code rate convolutional encoder and the same spectral efficiency with $\eta_{mGQSM} = \eta_{MA-SM} = 9$ bpcu.

and $N_r = 4$ system model for the MA-SM. We compare the BER performance of mGQSM and MA-SM with same number of RF-chains, $N_{rf} = 2$. For the probability of error 10^{-3} , we observe that the BER performance of mGQSM system with 4-QAM modulation scheme gain 1.5 dB value in SNR over MA-SM system with 8-QAM modulation scheme.

Fig. 4 shows the the BER performance of mGQSM and GQSM systems with the same system configuration, 4 by 4 system model with $N_t = 4$ and $N_r = 4$, number of RF chains, $N_{rf} = 2$ and 4-QAM modulation scheme. At probability

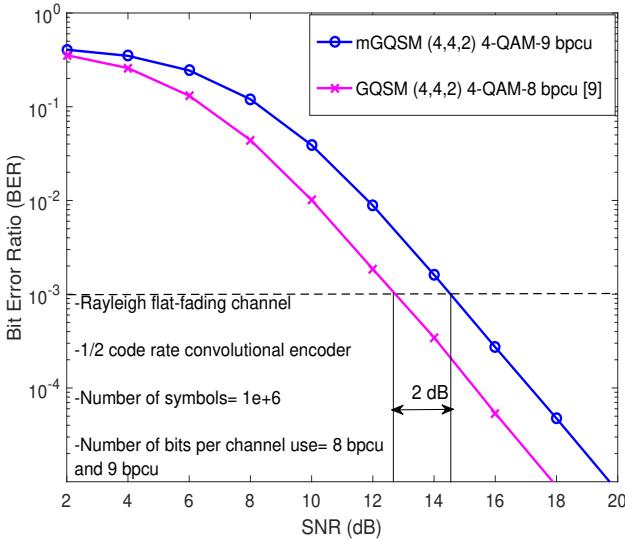


Figure 4. BER performance vs SNR for mGQSM and GQSM with $\frac{1}{2}$ code rate convolutional encoder with $N_t = 4$ and $N_r = 4$ system model and the spectral efficiencies with $\eta_{mGQSM} = 9$ bpcu and $\eta_{GQSM} = 8$ bpcu.

of error 10^{-3} , we observed that the BER performance of mGQSM system with spectral efficiency of 9 bpcu loose 2 dB value in SNR over GQSM system with spectral efficiency of 8 bpcu. However, the proposed mGQSM scheme provide extra one bpcu spectral efficiency over GQSM.

Let the 4 by 4 mGQSM system model with $N_t = 4$ and $N_r = 4$ and 4-QAM modulation scheme. In addition to the perfect CSIR, we use $N_r \times N_t$ error channel matrix with all the elements in the matrix are assumed to be i.i.d $\mathcal{CN}(0, \sigma_h)$. Fig. 5 shows the BER performance of the mGQSM system with perfect and imperfect CSIR conditions with the variance σ_h is varies from 0.01 to 0.05. For the probability of error 10^{-3} , we observe that imperfect CSIR with variance 0.01 and 0.02 are loses approximately 1.5 dB and 4.5 dB values in SNR respectively compared to the mGQSM with perfect CSIR. Therefore, we conclude that the BER performance is degrades by considering imperfect knowledge of the channel.

V. CONCLUSION

In this paper, we proposed a modified GQSM scheme without antenna grouping and we use multiple RF chains to enhance the spectral efficiency. The proposed scheme provides an additional one bpcu spectral efficiency over GQSM scheme. We study the performance via numerical simulations using half code rate convolutional encoder and Viterbi decoding algorithm to estimate the transmitted bits. We compute the computational complexity of ML-decoding in terms of real valued multiplications and introduce a variant of mGQSM called RC-mGQSM to reduce the complexity by decreasing the spectral efficiency. We also presents the BER performance of mGQSM system with imperfect channel conditions.

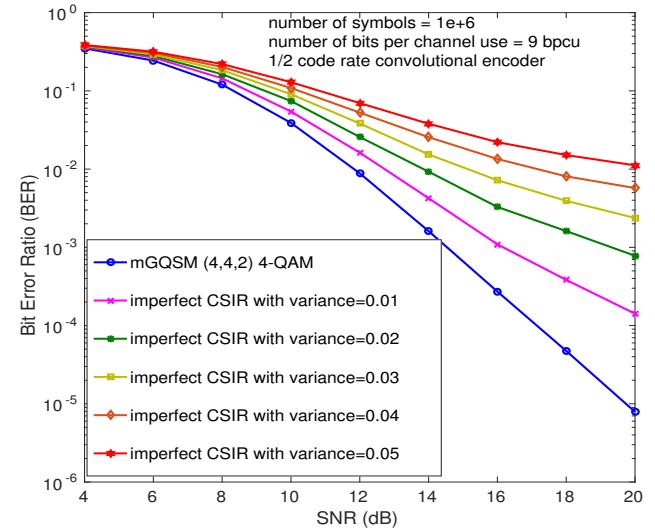


Figure 5. BER performance vs SNR for mGQSM ($N_t = 4$, $N_r = 4$, $M = 4$) system with perfect and imperfect CSIR with the error channel entries are i.i.d $\mathcal{CN}(0, \sigma_h)$ and the variance $\sigma_h = 0.01$ to 0.05

REFERENCES

- [1] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-blast: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *1998 URSI International Symposium on Signals, Systems, and Electronics. Conference Proceedings*, pp. 295–300, Sep 1998.
- [2] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 2228–2241, July 2008.
- [3] M. D. Renzo, H. Haas, A. Ghayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized mimo: Challenges, opportunities, and implementation," *Proceedings of the IEEE*, vol. 102, no. 1, pp. 56–103, 2014.
- [4] R. Rajashekhar, K. V. S. Hari, and L. Hanzo, "Antenna selection in spatial modulation systems," *IEEE Communications Letters*, vol. 17, pp. 521–524, March 2013.
- [5] J. Jeganathan, A. Ghayeb, L. Szczecinski, and A. Ceron, "Space shift keying modulation for mimo channels," *IEEE Transactions on Wireless Communications*, vol. 8, pp. 3692–3703, July 2009.
- [6] R. Mesleh, S. S. Ikki, and H. M. Aggoune, "Quadrature spatial modulation," *IEEE Transactions on Vehicular Technology*, vol. 64, pp. 2738–2742, June 2015.
- [7] A. Younis, N. Serafimovski, R. Mesleh, and H. Haas, "Generalised spatial modulation," in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, pp. 1498–1502, Nov 2010.
- [8] J. Wang, S. Jia, and J. Song, "Generalised spatial modulation system with multiple active transmit antennas and low complexity detection scheme," *IEEE Transactions on Wireless Communications*, vol. 11, pp. 1605–1615, April 2012.
- [9] F. R. Castillo-Soria, J. Cortez-González, R. Ramirez-Gutierrez, F. M. Maciel-Barboza, and L. Soriano-Equigua, "Generalized quadrature spatial modulation scheme using antenna grouping," *ETRI Journal*, vol. 39, no. 5, 2017.
- [10] T. L. Narasimhan, P. Raviteja, and A. Chockalingam, "Generalized spatial modulation in large-scale multiuser mimo systems," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 3764–3779, July 2015.

Predistortion Linearizer Design for K_u Band RF Power Amplifier

Girish Chandra Tripathi, Meenakshi Rawat

Department of Electronics and Communications Engineering

Indian Institute of Technology, Roorkee, Uttarakhand, India

Email: gtripathi@ec.iitr.ac.in, rawatfec@iitr.ac.in

Abstract—This paper presents a K_u band analog predistorter linearizer for improving the linearity of high-power radio frequency (RF) amplifiers. This method is applicable for both solid-state power amplifiers (SSPA) and traveling-wave tube amplifiers (TWTA). The designed linearizer consists of analog components, hence it is wideband as compared to the digital predistortion. Moreover, due to most of the passive components, the circuit is simpler and with the advantage of individual control of amplitude modulation to amplitude modulation (AM-AM) and amplitude modulation to phase modulation (AM-PM) conversions. For proof of concept, the designed linearizer is simulated with ZX60-14012L+ class AB SSPA. The S2D SSPA model is extracted using vector network analyzer. The proposed linearizer shows a reduction in third order intermodulation of 37 dB approximately at 4 dB output power backoff for two-tone signal. Similarly, for Long-Term Evolution (LTE) 20 MHz signal after linearization adjacent channel power ratio is approximately 47 dBc and shows a correction of 16 dB.

I. INTRODUCTION

The demand for high data rates is increasing day by day and we have moved to Long-term evolution (LTE) and Long-term evolution advanced (LTE-A) from the advanced mobile phone system. This quest for higher data rate is leading from 4th generation to 5th generation. These requirements are imposing new challenges on radio frequency (RF) front ends like efficiency, power consumption and non-linearity. These problems are mostly due to the RF power amplifiers (RF-PA) which degrade the overall performance of the RF transmitter. When RF-PA is operated near the saturation point, they are most effective in terms of delivered output power versus supplied power, however, it comes at the cost of strong nonlinear distortion. This distortion impacts the information bearing signal. Hence at the high input power, the RF-PA will show good efficiency at the cost of linearity. Therefore a trade-off exists between linearity and efficiency of RF-PA.

To achieve efficiency as well as linearity simultaneously, pre-distortion is a prominent method, which is used very frequently for RF-PA nonlinearity compensation [1], [2]. Moreover, these challenges become more prominent at high frequencies such as K_u , band which is defined by International Telecommunication Union (ITU) in the range of 12-18 GHz [3]. The predistortion further can be classified as analog predistortion (APD), digital predistortion (DPD) and Hybrid predistortion (HPD). Although APD has less accuracy

than the DPD, it is advantageous because (1) Analog to digital conversion is not required (2) Access to baseband signal is not required (3) APD has much smaller footprint and cost. Therefore, this paper focuses on APD due to its compactness and advantages of $RF_{in} - RF_{out}$ system. The APD linearization given in [4] uses a nonlinearity generator (NLG). The magnitude and phase at the output are adjusted using a vector multiplier (VM) and then combined to cancel the intermodulation distortion (IMD) components. Further, a linearizer is given in [5] where a conventional APD is used with a phase shaper block to additionally tune the phase of the signal at RF-PA input. This scheme uses multiple paths for nonlinearity compensation. In [6], APD is presented using Schottky diodes for third and fifth order IMD reduction which works significantly up to 6 GHz. In [7], [8], [9], a multi-branch APD is proposed which uses components like phase shifter, VM, attenuator, a delay element, hybrid coupler and envelope detector etc. for linearization. A dual Schottky diode based linearizer for the radio-over-fiber transmitter is presented in [10] and it is non-tunable after deployment. The [11], [12] provides diode based linearizer and its temperature compensation method. Moreover a reflection based linearizer is presented in [13], [14]. In [15], an E-Band APD and RF-PA Monolithic Microwave Integrated Circuit (MMIC) Chipset are presented which is very compact but the correction obtained is less. Further, in [16] a K_u band cascaded linearizer is presented for TWTA. The above-mentioned method uses a nonlinearity generator (NLG) circuit and a VM or attenuator phase shifter

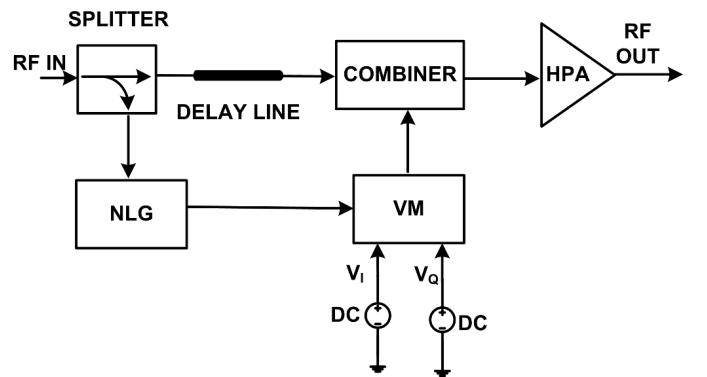


Fig. 1. Block diagram of the conventional analog predistorter circuit [4].

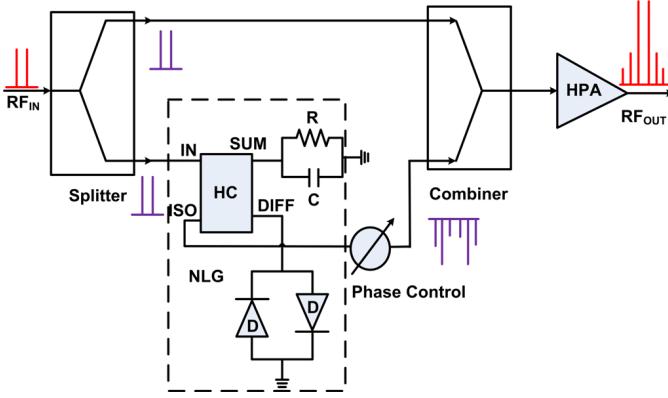


Fig. 2. Modified analog predistorter.

for linearizer design. The VM suffers from imperfections like in-phase (I) and quadrature (Q) imbalance. In addition the preamplifier used in VM may possibly go to saturation which again causes distortions. Most of the commercial VM are available in the frequency range upto 2.7 GHz [17], [18]. Some commercial VM are available [18] from the frequency range of 8 GHz to 16 GHz but they have disadvantage in terms of high noise factor which varies typically from 8 to 22 dB at different gain and temperature condition. The RF-delay line used in some of the cases above also increases the size of circuit and therefore these multiple components will require a large area and cause inherent analog imperfections, which will be further propagated in the overall system. Therefore to avoid these problems, all the components should be designed on a single board. This will help in reduction in the connector losses and an increase in the compactness. In this paper, we propose an APD linearizer on a single board and eliminate the requirement of time delay in the direct branch and VM in the nonlinear branch as mentioned in [4].

This paper is structured as follows: The predistorter circuit design is described in Section-II. The operating principle discussed in Section III. Section IV depicts the linearization results. The conclusion is drawn in section V.

II. CIRCUIT DESIGN

The configuration presented in [4] is modified and the VM has been replaced by a phase shifter to overcome the imperfections of the VM and the delay line is replaced by a transmission line. The modified architecture consists of the power splitter, power combiner, NLG, phase shifter, and Schottky diodes. The modified structure is shown in Fig. 2, and constitutes of following components.

A. Power divider/combiner

The divider is used to split the RF signal path into two paths. The power combiner is used to combine the signals from both paths and provided to the RF-PA. Fig. 3 shows the S-parameters of the splitter, where the return-loss is less than -15 dB and desired transmission for 6 GHz bandwidth is obtained at a center frequency of 13.75 GHz.

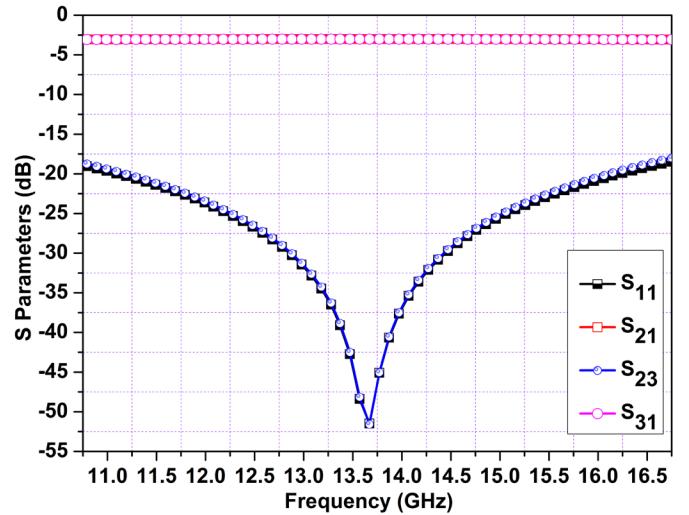


Fig. 3. S-Parameters of Power divider/combiner.

B. Nonlinearity Generator (NLG)

In Fig. 2, the lower branch shows the architecture of NLG (shown in the dotted box). The NLG consists of 180^0 hybrid coupler, diodes, and an RC network. The antiparallel Schottky diodes are used to generate the intermodulation (IM) components and are realized using HSCH-5314 diodes from Avago technologies. These diodes are capable of working from the frequency range of 1-26 GHz and can be operated in self-bias as well as in external bias mode. The combination of resistor and capacitor are used to cancel the fundamental signal and the diode network for nonlinearity generation. Therefore, the output of NLG consists of IM components along with the attenuated main tone. The Fig. 4 shows the S-parameter of the coupler, where the obtained return loss is less than -15 dB and desired transmission for 6 GHz bandwidth obtained at a center frequency of 13.75 GHz. A phase shifter is used at the

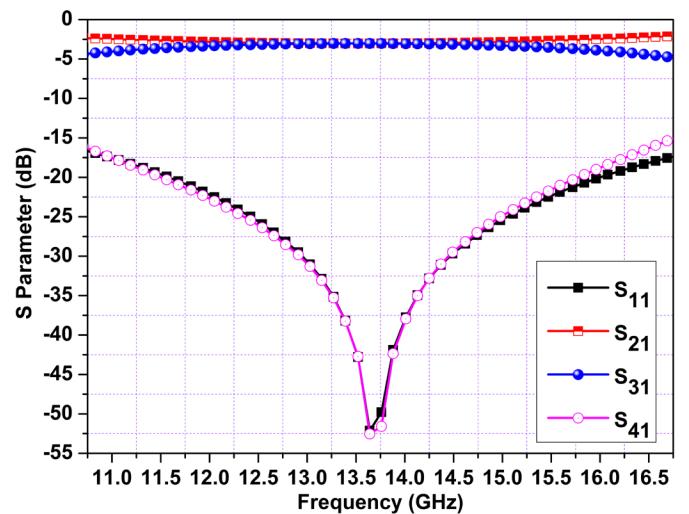


Fig. 4. S-parameter of the hybrid coupler.



Fig. 5. Phase shift vs. Frequency response of the phase shifter

output of NLG to tune the phase of IM components generated by NLG which are opposite phase of the IM components generated by RF-PA so that they can cancel out each other. The phase is controlled using a reflection type phase shifter which is designed by four port coupler with variable load. We have used SMV2019 Varactor diodes as a variable load to obtain different phase shift. This phase variation is obtained by changing the voltage applied to the diode. The Fig. 5 shows the simulated phase shift response of the phase shifter having wide phase shift ranging from 11 to 16 GHz.

C. RF-PA and Signal

A two-tone signal at frequencies 13.745 GHz and 13.755 GHz with a spacing of 10 MHz is used. Moreover, to test the performance with wideband signal we have also taken LTE 20 and 100 MHz signal at a sampling rate of 92.16 MSPS (upsampled from 30.72 MSPS). We have used RF-PA ZX60-

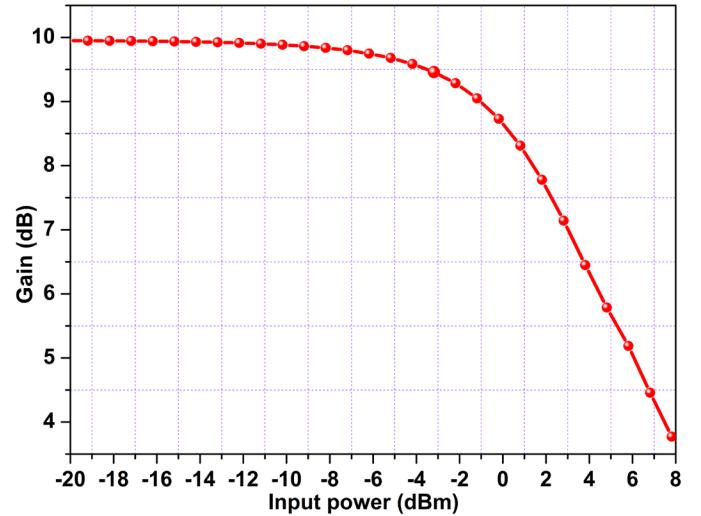


Fig. 7. AM-AM characteristics of SSPA.

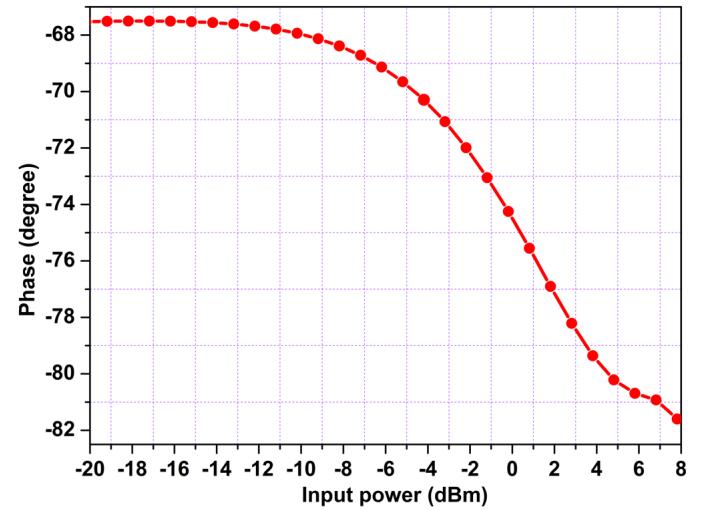


Fig. 8. AM-PM characteristics of SSPA.



Fig. 6. Measurement setup for s2d file design.

14012L+ from minicircuit. The operating frequency of this power amplifier is from 300 KHz to 14 GHz and the typical gain is 12 dB.

D. S2D Design for RF-PA

The S-parameter data (.s2d) file is created using the real measurement obtained from RF-PA. The objective of this measurement is to translate the behavior of real RF-PA into a simulation environment for analysis. Moreover, the linearizer created using this design will be able to match the real condition of RF-PA. This file consists of (a) the small signal measurement as a function of frequency, (b) Non-linear characteristics of the amplifier. For the gain compression and the phase compression, we have measured AM-AM and AM-PM curves with vector network analyzer as a function of input power level. The measurement setup for s2d file design is shown in Fig. 6, where Keysight PNA-X is used for frequency

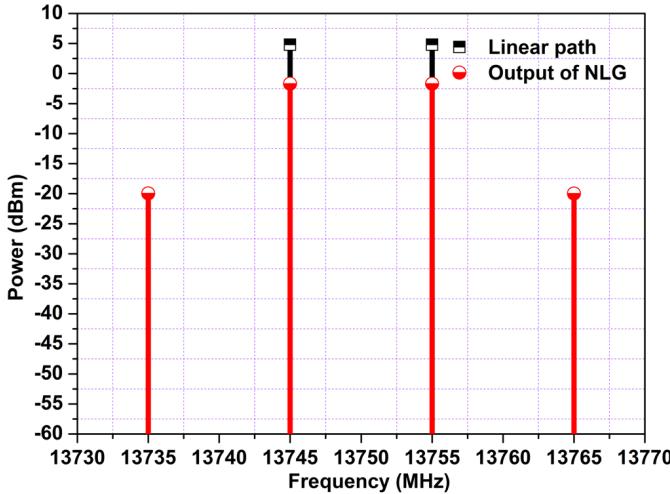


Fig. 9. Output spectrum of Direct and nonlinear paths for two-tone signal.

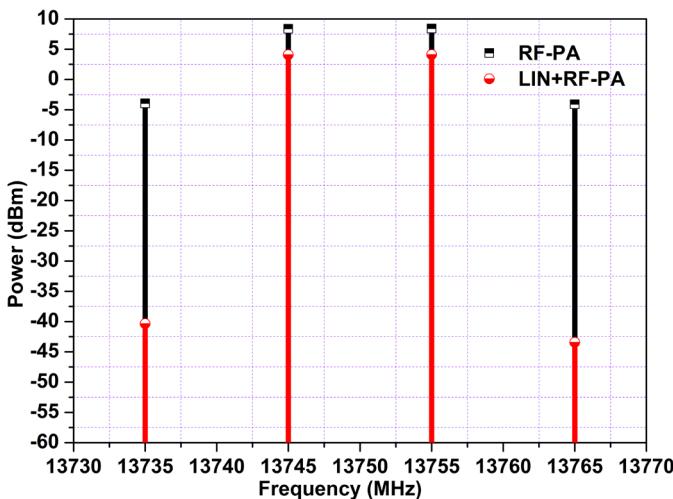


Fig. 10. Output spectrum of SSPA and Linearizer+SSPA for the two-tone signal.

and power sweep measurement. Fig. 7 shows the AM-AM characteristics of the RF-PA plotted using s2d file which shows a gain compression of approximately 5 dB. Similarly, in Fig. 8, AM-PM characteristics shows a phase compression of 14° .

III. WORKING PRINCIPLE

The working principle can be better understood with the help of Fig. 2. This design is an $RF_{in} - RF_{out}$ system which did not require baseband access for linearization. Firstly the signal is passed through 1:2 splitter, the output upper branch is a direct linear path which introduces some delay only. The lower path is the nonlinear path which consists of NLG for the distortion generation. The distortions are basically IM components and have less power as compared to the main tone signals. They are passed through a phase shifter, which generates the inverse phase corresponding to the IM output of RF-PA and results in the IM cancellation. For analysis, the phase shifter available in the simulator is used, it can also

be designed using varactor diodes and reflective networks with tuning capability. Since the transmission line used in the upper branch will provide some delay and the additional delay can be compensated using phase shifter as they are interrelated [19]. Fig. 9 shows the output spectrum signal from the direct branch linear path and NLG output. It can be seen that the main tone in the linear path is higher while in the output of NLG it is lower. Similarly, there are 3IM components at the output of NLG due to distortion generation. The signal from both paths are combined through a 2:1 combiner and the output is passed through RF-PA and the combined output is linear.

IV. RESULTS

To analyze the advantage of APD at such high frequencies an SSPA is simulated using advanced design system software whose behavior is shown in Fig. 7 and Fig. 8. A two-tone signal at frequency 13745 and 13755 MHz with is used for simulation. Fig. 10 shows the output spectrum of without (RF-PA output) linearization and with linearization (Linearizer+RF-PA) which shows a cancellation of 36.42 dB in $3IM_{left}$ at 4 dB output power backoff (OPBO) which adheres to the limits defined in [20]. Similarly 39.31 dB cancellation in $3IM_{right}$ is achieved at same OPBO. To test the performance of the linearization scheme, it is tested with LTE 20 MHz signal as shown in Fig. 11. The results show there is a correction of 16 dB in adjacent channel power ratio and it is found to be 47 dBc at 2 dB OPBO which qualifies the mask. The same trend is followed when tested with LTE 100 MHz signal as shown in Fig. 12. The Fig. 13 depicts the C/IM performance without linearizer (RF-PA) and with linearizer (LIN+RF-PA). Moreover, the performance is plotted in terms of OPBO with SSPA as well as linearizer for a fair comparison with two-tone signal. The Fig. 13 shows the carrier to 3^{rd} IM ratio of RF-PA output with and without linearizer shows that by using LIN+RF-PA the C/IM ratio is increasing with OPBO.

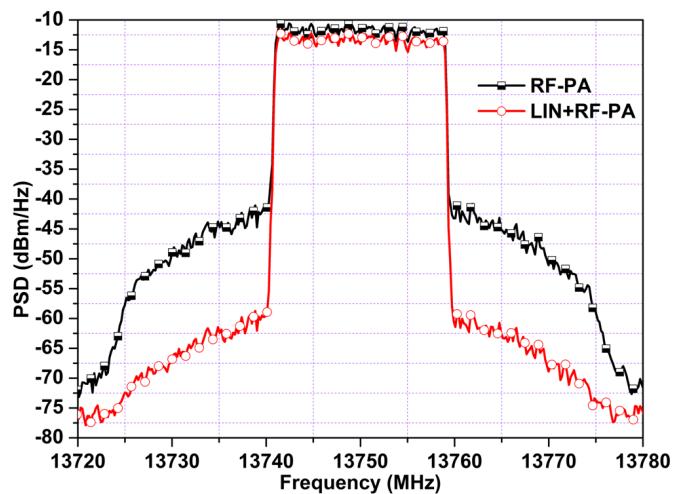


Fig. 11. Frequency spectrum of SSPA and Linearizer+SSPA for LTE 20 MHz bandwidth signal.

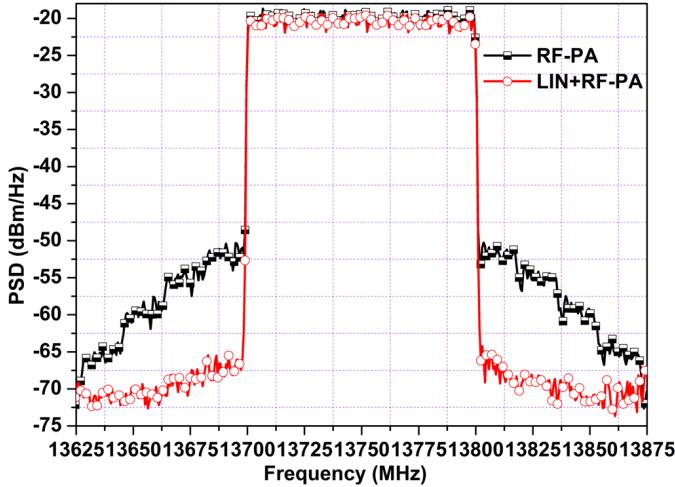


Fig. 12. Frequency spectrum of SSPA and Linearizer+SSPA for LTE 100 MHz bandwidth signal.

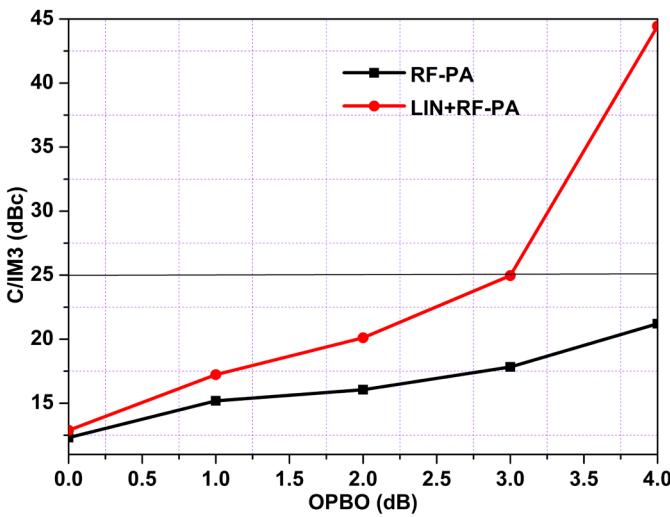


Fig. 13. C/IM3 performance of SSPA and Linearizer+SSPA for two-tone signal.

The C/IM ratio is found to be 44.44 dBc as compared to 21.21 dBc with SSPA at 4 dB OPBO.

V. CONCLUSION

In this paper, a linearizer for K_u band frequency is presented. Simulation results show that there is a good correction in IM3 with a two-tone signal test of 36-39 dB and 16 dB correction for LTE 20 MHz signal. This method has the capability of wideband linearization (LTE 100 MHz), as the designed splitter and hybrid coupler are very wideband. There is no requirement of VM which was the main band limiting component. Moreover by increasing the bandwidth of the phase shifter a more wideband linearizer can be achieved. This linearizer is useful for satellite communication as it is compact in nature. The proposed linearizer can be used for repeater systems where only RF signal is available.

ACKNOWLEDGMENT

This work was supported in part by Defense Research and Development Organization, India, under Grant ERIP/ER/1400477/M/01 and in part by the Science and Engineering Research Board, India, under Grant EMR/2016/001310. The authors would like to thank A. K. Sahoo for his suggestion.

REFERENCES

- [1] J. X. Qiu, D. K. Abe, T. M. Antonsen, B. G. Danly, B. Levush, and R. E. Myers, "Linearizability of TWAs using predistortion techniques," *IEEE Trans. Electron Devices*, vol. 52, no. 5, pp. 718-727, May 2005.
- [2] A. Katz, J. Wood, and D. Chokola, "The evolution of PA linearization: From classic feedforward and feedback through analog and digital predistortion," *IEEE Microw. Mag.*, vol. 17, no. 2, pp. 32-40, Feb. 2016.
- [3] International telecommunication union "Nomenclature of the frequency and wavelength bands used in telecommunications," *Recommendation ITU-R V.431-8* (08/2015).
- [4] G. C. Tripathi and M. Rawat, "RF_{in}-RF_{out} Linearizer System Design for Satellite Communication," *IEEE Trans. Electron Devices*, vol. 65, no. 6, pp. 2378-2384, June 2018.
- [5] B. Shi, S. W. Leong, B. Luo and W. Wang, "A novel GHz bandwidth RF predistortion linearizer for Ka band power amplifier," *IEEE Region 10 Conf. (TENCON-2017)*, Penang, 2017, pp. 1610-1613.
- [6] X. Zhang, S. Saha, R. Zhu, T. Liu, and D. Shen, "Analog pre-distortion circuit for radio over fiber transmission," *IEEE Photon. Technol. Lett.*, vol. 28, no. 22, pp. 2541-2544, Nov. 2016.
- [7] Y. S. Lee, M. Y. Lee, and Y. H. Jeong, "A wideband analog predistortion power amplifier with multi-branch nonlinear path for memory-effect compensation," *IEEE Microw. Wireless Compo. Lett.*, vol. 19, no. 7, pp. 476-478, Jul. 2009.
- [8] K. J. Cho, D. H. Jang, S. H. Kim, J. Y. Kim, J. H. Kim, and S. P. Stapleton, "An analog compensation method for asymmetric IMD characteristics of power amplifier," *IEEE Microw. Wireless Comp. Lett.*, vol. 14, no. 4, pp. 153-155, Apr. 2004.
- [9] J. Yi, Y. Yang, M. Park, W. Kang, and B. Kim, "Analog predistortion linearizer for high power RF amplifier," *IEEE MTT-S Int. Microwave Symp. Dig.*, vol. 3, Boston, MA, June 2000, pp. 1511-1514.
- [10] R. Zhu, X. Zhang, D. Shen, and Y. Zhang, "Ultra broadband pre-distortion circuit for radio-over-fiber transmission systems," *J. Lightw. Technol.*, vol. 34, no. 22, pp. 5137-5145, Nov. 15, 2016.
- [11] S. C. Bera, R. V. Singh, and V. K. Grag, "Diode-based predistortion linearizer for power amplifier," *IET Electron. Lett.*, vol. 44, no. 2, pp. 125-126, Jun. 2008.
- [12] S. C. Bera, V. Kumar, S. Singh, and D. K. Das, "Temperature behavior and compensation of diode-based pre-distortion linearizer," *IEEE Microw. Wireless Compon. Lett.*, vol. 23, no. 4, pp. 211-213, Apr. 2013.
- [13] L. Jie, Z. Hua-Dong and L. Zeng-liang, "A novel two-branch predistortion linearizer of Ku-band TWT in communication applications," *IET Int. Radar Conf.*, Hangzhou, 2015, pp. 1-3.
- [14] H.-L. Deng, D.-W. Zhang, D.-F. Zhou, S.-X. Wang, and J.-X. Wang, "A Ka-band analog predistortion linearizer for travelling wave tube amplifiers," *IEEE Asia-Pacific Microw. Conf.(APMC)*, Nanjing, 2015, pp. 1-3.
- [15] M. Gavell, G. Granström, C. Fager, S. E. Gunnarsson, M. Ferndahl and H. Zirath, "An E-Band Analog Predistorter and Power Amplifier MMIC Chipset," *IEEE Microw. Wireless Compon. Lett.*, vol. 28, no. 1, pp. 31-33, Jan. 2018.
- [16] C. Mallet, C. Duvanaud, L. Carr and S. Bachir, "Analog predistortion for high power amplifier over the Ku-band (13,7514,5 GHz)," *IEEE European Microw. Conf. (EuMC)*, Nuremberg, 2017, pp. 848-851.
- [17] Maxim Integrated, "Vector Modulators / Multipliers," Accessed: Sept. 25, 2018, [Online] Available: <https://para.maximintegrated.com/en/search.mvp?fam=vector&tree=wireless>
- [18] Analog Devices, "Vector-modulators," Accessed: Sept. 25, 2018, [Online] Available: <http://www.analog.com/en/products/rf-microwave/phase-shifters-vector-modulators/vector-modulators.html>
- [19] G. C. Tripathi and M. Rawat, "Delay compensation for 4G/5G transmitter system characterization," *Microw., Opt. Technol. Lett.*, vol. 59, no. 8, pp. 1887-1890, 2017.

- [20] MIL-STD-188/164C, Department of defense interface standard: interoperability of SHF satellite communications terminals, [Online] Available: http://everyspec.com/MIL-STD/MIL-STD-0100-0299/MIL-STD-188-164C_55855.

LSTM-Deep Neural Networks based Predistortion Linearizer for High Power Amplifiers

Deepmala Phartiyal

Indian Institute of Technology Roorkee
Roorkee – 247667, INDIA
trideepa@gmail.com

Meenakshi Rawat,

Department of Electronics and

Communication Engineering

Indian Institute of Technology Roorkee
rawatfec@iitr.ac.in

Abstract—Linear high power amplifiers (HPAs) are the need of current communications technology. But, almost all PAs show non-linear characteristics during amplification which are reflected in the transmitted signal in the form of distortions. Linearization is a process to suppress the effect of the non-linear characteristic of a PA. Various methods are available to perform linearization. Predistortion (PD) linearization methods are very successful due to its simplicity in design and ease of integration with PAs. PD linearization methods observe the PA dynamic characteristics (nonlinearity) and then formulate an “inverse transfer function” to suppress this non-linearity. In the last decade, machine learning (ML) based PD linearizers are proposed and proved useful. Since then, numerous ML-PD linearizers have been developed. Shallow neural networks (NNs) based PD linearizers are successfully used to map the inverse transfer function but lack generalization performance in the presence of system conditions (IQ imbalance, DC offset). With deep learning (DL) technology, deep neural networks (DNNs) can map the complex inverse transfer function under different system conditions. This study proposes a long short-term memory (LSTM) DNN based PD linearizer for linearization of PA under different conditions. In this study, it is shown that LSTM is able to extract and exploit memory effects of PA over the perceptron. Comparative results with shallow NNs suggest reliable potential in the proposed DNN model in terms of generalization performance.

Keywords: *Power Amplifiers, Nonlinearity, Predistortion, DNN, LSTM*

I. INTRODUCTION

High power amplifiers (HPAs) are used in all sorts of communication services such as in satellite communications or terrestrial microwave communication to transmit multiple signals in large quantities at high data rates over long distances. Klystron and travelling wave tube power amplifiers (KPAs, and TWTPAs) are more technoeconomical than solid state PAs (SSPAs) in the microwave region [1]. One major shortcoming with KPAs and TWTPAs is their non-linear characteristics during amplification process. If this non-linearity is compensated by some linearization approach, then KPAs and TWTPAs can achieve better performances. Linearization is a systematic process of reducing the distortions caused by non-linearity. Numerous linearization techniques have been developed so far. Usually, linearizing components are assembled together and boxed under a label called ‘Linearizer’. Most popular categories of linearizers are; feedforward, feedback, and predistortion. Feedforward linearization approaches work well with TWT PAs and KPAs but are most widely used with SSPAs. These

approaches are rather complex to implement and integrate with the PA. Feedback approaches work well at RF, but less work is documented at microwave. These approaches suffer from issues like amplifier stability, or robustness over wide frequency bands [1].

The working principle of a PD linearizer is explained in Section II. A comparative study of different PD linearizers over performance with mildly nonlinear to highly nonlinear PAs is given in [2]. The study suggests that machine learning (ML) based linearizers are better suited for the application. Although the design parameters are relatively high, but ML linearizers are more popular recently [3][4][5]. The major issue with shallow neural networks (NN) PD linearizers was its low generalization capability. With the introduction of “Deep Learning (DL)” technology, complex function approximations and classifications have been much reliable and accurate [6]. PD linearizer modelling with DL may have the potential to capture and exploit memory effects of PAs because of deeper layers. It is also established now that DL provides balance between accuracy and generalization in models under different data conditions [7].

In this study, a long short term memory (LSTM) deep neural network (DNN), i.e. LSTM-DNN is proposed as a PD linearizer for linearization of PA nonlinearity under different system conditions such as DC offset or IQ imbalance. LSTMs are recurrent neural networks (RNNs) which are used for time series data modelling and prediction. LSTM units are used to capture and exploit the memory effects of PA. It is shown in this study that the proposed LSTM-DNN PD linearizer is advantageous over shallow NN linearizers when applied to Doherty PA under different conditions.

The article is divided as follows. The first section provides problem statement and history of various PA linearizers. Section II illustrates theoretical background and key concepts required in the study. This section also introduces the proposed PD model and describes its architecture, training and implementation. Section III provides data information used in this study, experiment setup, and results are also discussed in this section.

II. LSTM-DNN PD LINEARIZER

A. Predistortion (PD) Linearizer: Concept

Predistortion (PD) approaches are popular in terrestrial microwave and satellite applications. These approaches are simpler to implement and easy to integrate with the PA. PD linearizers generate a non-linear transfer function which is

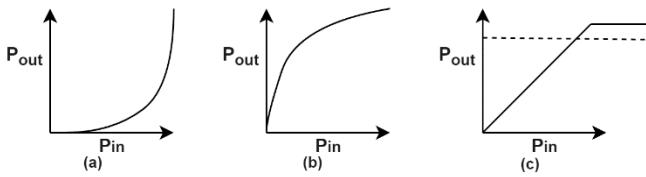


Fig. 1. Working principle of a PD linearizer. (a) PD characteristic, (b) PA characteristic and, (c) Linearized output.

inverse to the transfer function of PA. The working principle of a PD linearizer is shown in figure 1. It is clear from figure 1(c) that up until a certain output power, the HPA output is linear in the presence of a PD linearizer. In the operating range, the PD linearizer generates intermodulation distortions (IMDs) of its own (equal in magnitude but 180 degrees out of phase with the IMDs of the PA). So, at the output, these IMDs cancel other and linear amplified power is achieved.

B. LSTM Unit

LSTM units are units of a recurrent neural network (RNN). A network of such units is called an LSTM network [8]. A common peephole LSTM unit is composed of a cell, an input gate, an output gate and a forget gate as shown in figure 2 [9]. In figure 2, each of these gates can be thought as a "standard" neuron of a feed-forward (or multi-layered) NN, that is, all the gates compute an activation (using an activation function) of a weighted sum ($Wx+b$) similar to standard neurons. In figure 2, i_t , o_t , and f_t represent the activations of respectively the input, output and forget gates, at time step t . The three exit arrows from the memory cell c to the three gates i , o , and f represent the peephole connections [9]. These peephole connections actually denote the contributions of the activation of the memory cell c at time step $(t-1)$, i.e. the contribution of c_{t-1} (and not c_t , as the picture may suggest). In other words, the gates i , o , and f calculate their activations at time step t (i.e., respectively, i_t , o_t , and f_t) also considering the activation of the memory cell c , at time step $(t-1)$, i.e. c_{t-1} . The single left-to-right arrow exiting the memory cell is not a peephole connection and denotes c_t . The little circles containing a \otimes symbol represent an element-wise multiplication between its inputs. The big circles containing an S-like curve represent the application of a differentiable function (like the sigmoid function) to a weighted sum. In summary, with the help of memory cell and the gates an LSTM network is able to memorize weights over longer time steps. The application of such an arrangement helps in predicting time series phenomena where outputs depend on relatively older inputs. This property is utilized in this study as PAs also show memory effects.

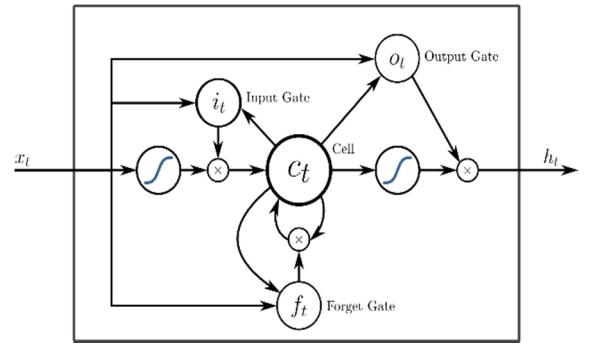


Fig 2. Architecture of a peephole LSTM cell with input (i.e. i), output (i.e. o), and forget (i.e. f) gates.

C. LSTM-DNN PD Linearizer

The proposed LSTM DNN model is motivated from two NN models developed for PA characteristic modelling. In the first [10], recurrent neural networks (RNNs) are utilized for dynamic behavioral modelling of PAs. The model uses a single hidden layer hence unable to efficiently map the memory effects of PA. The second [11], proposes a time delayed real valued shallow NN for PA behavioral modelling. Inspired by the above models, the current model is proposed. The proposed PD linearizer is shown in figure 3. Input layer consists of current and delayed inputs. The length of delay is presumed (let say m). Similar delay length is set for both I_{in} , and Q_{in} (I and Q are the in-phase and quadrature-phase components of a baseband signal respectively). The LSTM layer receives the signal sequence (current + delayed), processes it and provides corresponding output. Hyperbolic tangent and softmax are used as activation functions in this layer [12]. The outputs from LSTM layer fed into the fully connected (FC) layer. The FC layer is a traditional multi-perceptron layer and hence functions as usual. Rectified linear unit i.e. $ReLU$ is used as activation in this layer [12]. Finally, the output layer with two neurons provides the desired outputs i.e. I_{out} and Q_{out} . Linear activation function is used in this layer [12]. This activation function sums up the outputs of previous FC layer neurons and linearly maps them at the output

The outputs I_{out} and Q_{out} , are formulated at any instant of training (let say n) as given in equation 1 and 2.

$$I_{out}(n) = f_1(X_{in}(n)) \quad (1)$$

$$Q_{out}(n) = f_2(X_{in}(n)) \quad (2)$$

Where X_{in} is the total input which includes present and past inputs and past outputs. f_1 and f_2 are unknown functions approximated by the proposed LSTM-DNN model.

$$\begin{aligned} X_{in}(n) &= [I_{in}(n), I_{in}(n-1) \dots I_{in}(n-m), Q_{in}(n), \\ &Q_{in}(n-1) \dots Q_{in}(n-m), I_{out}(n), I_{out}(n-1), \dots \\ &I_{out}(n-t), Q_{out}(n), Q_{out}(n-1) \dots Q_{out}(n-t)] \end{aligned} \quad (3)$$

where, m is memory depth in input, and t is memory delay length in output (feedback signal).

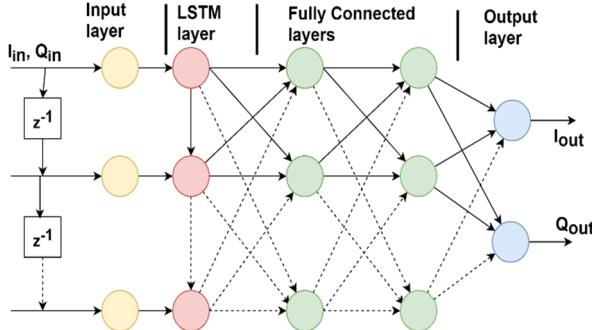


Fig 3. Proposed LSTM-DNN PD linearizer architecture

III. EXPERIMENT, RESULTS, AND DISCUSSION

A. Experiment Setup:

The training and testing of the proposed model is done using a realistic data. Details of PA and ancillary information used during the experiment is provided in table 1. PA dynamic characteristic data is obtained under different system conditions, a summary of data collected under different conditions is provided in table 2. In table 2, presence of IQ imbalance and/or DC offset distortions in the signal are marked as (✓) and their absence is marked as (X) symbols. Corresponding Gain/Phase vs input power (P_{in}) plots are shown in figure 4. Figure 4 depict the effects of IQ imbalance and/or DC offset distortions on the baseband signal. Phase vs Input Power plot of case a (see figure 4(b)) shows maximum distortion because it is affected by both IQ imbalance and DC offset distortions.

TABLE 1 TRAINING DATA CHARACTERISTICS

Parameter	Description
PA	54-dBm saturation power Doherty PA
SSG	Small Signal Gain = 38 dB
Input signal	Two carrier WiMAX signal
PAPR	Peak to Average Power Ratio - 10.5 dB

B. Model Setup and Training:

Model formulation, development, training, and testing is carried out in a Python environment with the help of ‘Keras’ deep learning library. Training is carried out in “batch” mode for faster weight update, supervised with the “backpropagation through time” training algorithm [13]. Mean squared error is taken as cost/loss function. Regularization during training to avoid overfitting is achieved using the “Dropout” method.[14].. ‘Adam’ stochastic optimizer is used for parameter optimization process [15]. Hyperparameter tuning is performed by “trial and error” method. Post the trial and error method based hyperparameter tuning, 10, and (7, and 5) are the number of neurons selected for the LSTM layer, and the two fully connected layers respectively.

The formulated model is trained with data from all cases i.e. case a, b, c, and d separately and individually (standalone cases). Also, model is trained with “combined” data which means data from all the cases are put together for training. Model is trained on combined data to achieve a generalized model which will act as PD for all cases of signal distortions.

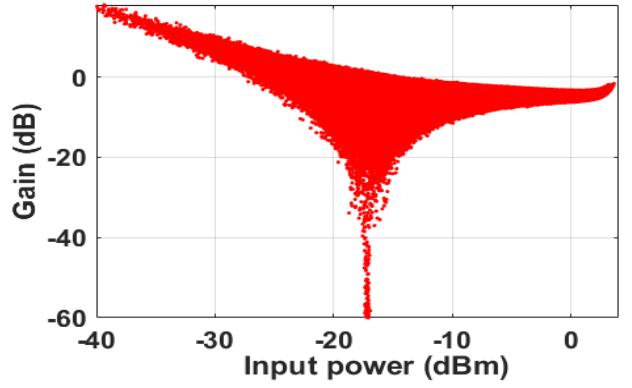
TABLE 2 SUMMARY OF PA DATA COLLECTED UNDER DIFFERENT CONDITIONS

Case No	Wireless standard	Signal Bandwidth (MHz)	IQ imbalance	DC offset	Data Samples
Case a	WiMAX	10	✓	✓	433125
Case b	WiMAX	10	✓	X	131049
Case c	WiMAX	10	X	X	51142
Case d	WiMAX	10	X	✓	51142

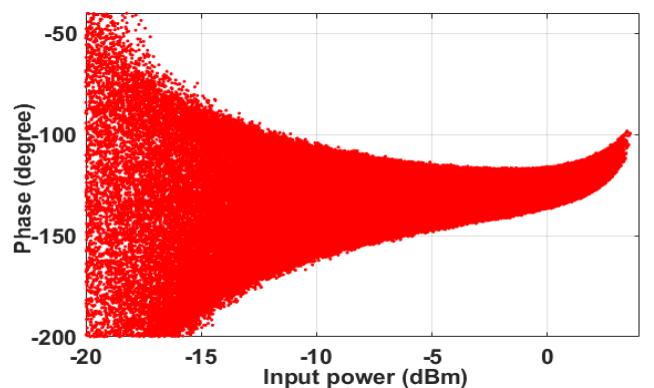
TABLE 3 SUMMARY OF NMSE RESULTS OBTAINED USING THE PROPOSED LSTM-DNN MODEL AND RVFTDNN MODEL

Case	RVFTDNN				LSTM-DNN			
	No of neurons in hidden layers (FC ₁ , FC ₂)*	No of epochs	NMSE (dB)	Memory depth	No of neurons in hidden layers (LSTM, FC ₁ , FC ₂)*	No of epochs	NMSE (dB)	Memory depth
Case a	7,15	500	-43.8353	1	10,7,5	500	-35.24	4
Case b	7,15	500	-26.1304	1	10,7,5	500	-24.24	4
Case c	7,15	500	-47.5481	1	10,7,5	500	-38.56	4
Case d	7,15	500	-25.43	1	10,7,5	500	-24.85	4
Combined	7,15	500	-21.586	1	10,7,5	500	-39.40	4

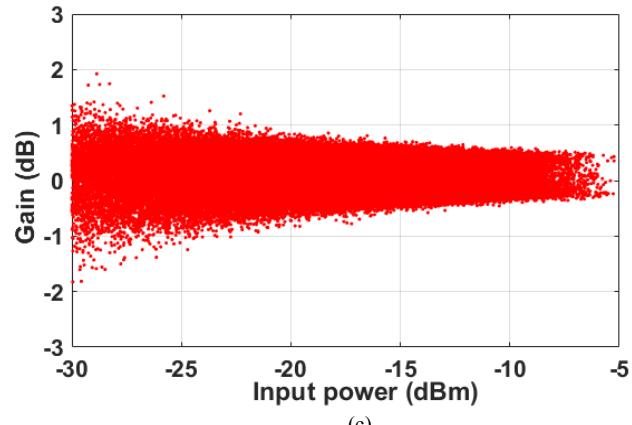
*LSTM = Long Short-Term Memory, FC = Fully Connected



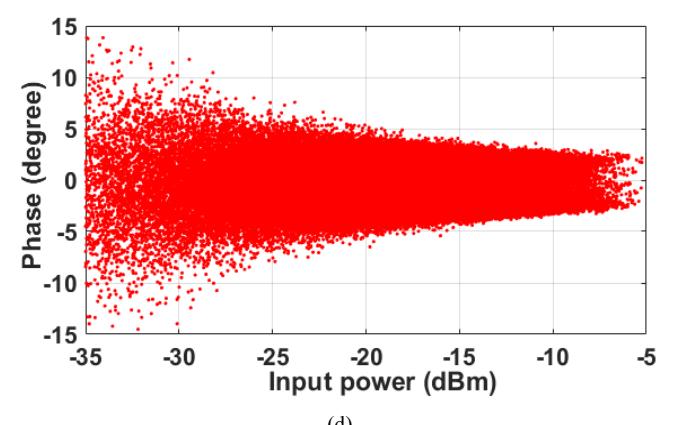
(a)



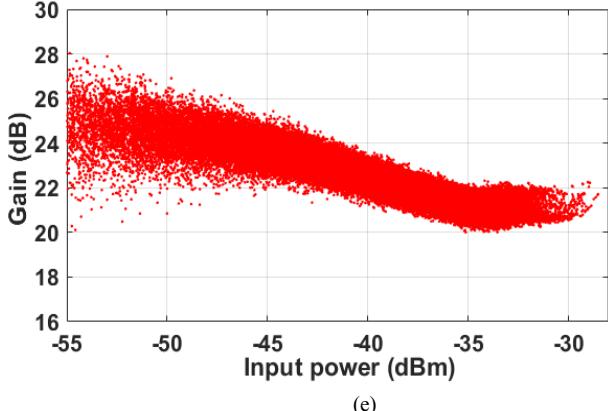
(b)



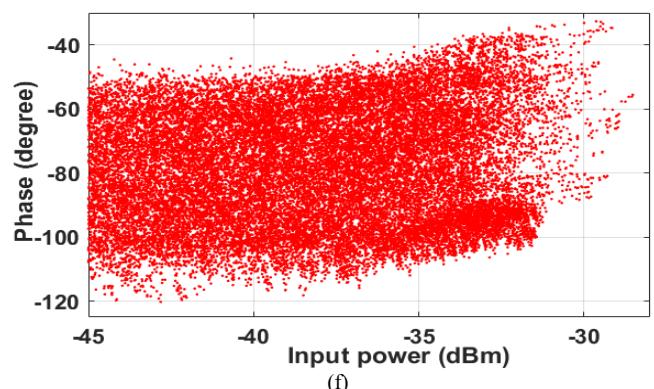
(c)



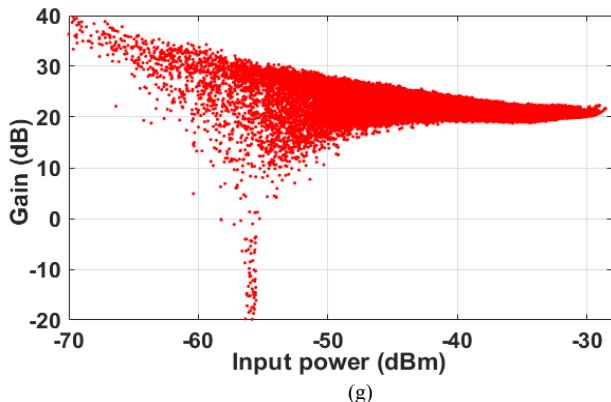
(d)



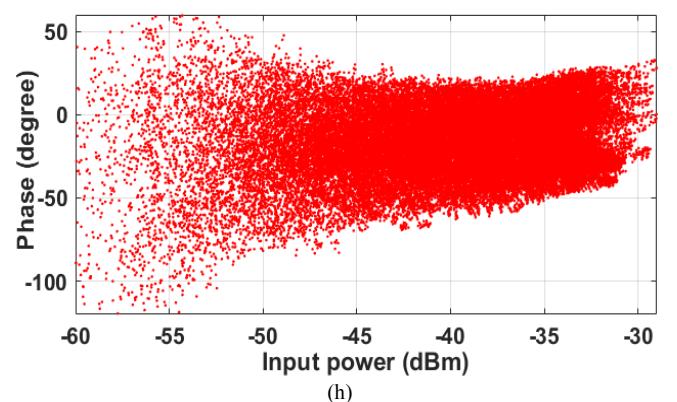
(e)



(f)



(g)



(h)

Fig 4. Gain/Phase vs Input Power plots for different conditions (mentioned in table 2). Figure (a) and (b) for Case a, data with both IQ imbalance and DC offset distortions. Figure (c) and (d) for Case b, data with IQ imbalance distortions only. Figure (e) and (f) for Case c, data with no distortions. Figure (g) and (h) for Case d, data with DC offset distortions only.

Performance evaluation is done based on the normalized mean square error (NMSE). The achieved NMSE is compared with earlier state of the art NN (RVFTDNN) model proposed in [3]. The comparison results are shown in table 3. From table 3, it is evident that RVFTDNN is better than LSTM-DNN in standalone cases (case a to case d). This is because of the universal approximation nature of the multilayer perceptron (MLP) over a given data. LSTM-DNN with NMSE value of -39 dB has shown better performance than RVFTDNN (-21 dB) for the combined data scenario. This is because of the memorizing nature of LSTMs and generalization capability of DNNs. Different value of memory depth for RVFTDNN and LSTM-DNN is based on the best performance of either model. With huge higher dimensional data, DNNs generalize better than shallow NNs. From table 3, it is clear that LSTM-DNN model generalizes well over multiple data scenario than shallow networks like RVFTDNN.

CONCLUSION

In this study, a LSTM-DNN based PD linearizer is proposed for linearizing PA non-linearity under different conditions. PA dynamic characteristics are collected in the presence of IQ impairments such as IQ imbalance and/or dc offset. A three hidden layer (one LSTM and two FC layer) is proposed for mapping PA characteristics. NMSE results obtained using the proposed model is compared with RVFTDNN model. Results show that the proposed model generalizes better than RVFTDNN model. Although the individual performance of proposed model is relatively low compared to RVFTDNN model results. The generalized LSTM-DNN PD model may be extended to map multiple PAs simultaneously in future.

ACKNOWLEDGMENT

Authors are thankful, to software defined radio (SDR) lab, IIT Roorkee, Roorkee for their resources and support.

REFERENCES

- [1] A. Katz, Linearization: “reducing distortion in power amplifiers,” *IEEE Microw. Mag.*, vol. 2, no. 4, pp. 37–49, 2001.
- [2] M. Rawat and F. M. Ghannouchi, ‘Distributed spatiotemporal neural network for nonlinear dynamic transmitter modeling and adaptive digital predistortion’, *IEEE Trans. Instrum. Meas.*, vol. 61, no. 3, pp. 595–608, 2012.
- [3] M. Rawat, K. Rawat, and F. M. Ghannouchi, ‘Adaptive Digital Predistortion of Wireless Power Amplifiers/Transmitters Using Dynamic Real-Valued Focused Time-Delay Line Neural Networks’, *IEEE Trans. Microw. Theory Tech.*, vol. 58, no. 1, pp. 95–104, Jan. 2010.
- [4] F. Mkadem and S. Boumaiza, ‘Physically Inspired Neural Network Model for RF Power Amplifier Behavioral Modeling and Digital Predistortion’, *IEEE Trans. Microw. Theory Tech.*, vol. 59, no. 4, pp. 913–923, Apr. 2011.
- [5] K. Li, N. Guan, and H. Wang, ‘Iterative Learning Control Assisted Neural Network for Digital Predistortion of MIMO Power Amplifier’, in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–5.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, ‘Deep learning’, *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
- [7] K. Kawaguchi, L. P. Kaelbling, and Y. Bengio, ‘Generalization in Deep Learning’, *arXiv preprint arXiv:1710.05468*, Oct. 2017.
- [8] Hochreiter Sepp and Schmidhuber Jürgen, ‘Long Short-Term Memory’, *MIT Press Journals*, vol. 9, no. 8, pp. 735–1780, 1997.
- [9] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, ‘Learning precise timing with LSTM recurrent networks’, *J. Mach. Learn. Res.*, vol. 3, no. 1, pp. 115–143, 2002.
- [10] D. Luongvih and Y. K.- Digest, ‘Behavioral modeling of power amplifiers using fully recurrent neural networks’, *IEEE MTT-S Int. Microw. Symp. Dig.*, pp. 1979–1982, Jun. 2005.
- [11] T. Liu, S. Boumaiza, and F. M. Ghannouchi, ‘Dynamic Behavioral Modeling of 3G Power Amplifiers Using Real-Valued Time-Delay Neural Networks’, *IEEE Trans. Microw. Theory Tech.*, vol. 52, no. 3, pp. 1025–1033, Mar. 2004.
- [12] Keras, ‘Activations - Keras Documentation’, *Keras*. [Online]. Available: <https://keras.io/activations/>. [Accessed: 02-Oct-2018].
- [13] Britz Denny, ‘Recurrent Neural Networks Tutorial, Part 3 – Backpropagation Through Time and Vanishing Gradients’, *WILDML*, 2015. [Online]. Available: <http://www.wildml.com/2015/10/recurrent-neural-networks-tutorial-part-3-backpropagation-through-time-and-vanishing-gradients/>. [Accessed: 25-Dec-2018].
- [14] N. Srivastava, G. Hinton, and A. Krizhevsky, ‘Dropout: a simple way to prevent neural networks from overfitting’, *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [15] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, *arXiv preprint arXiv:1412.6980* (2014). Dec. 2014.

Design of Multiband Negative Permittivity Metamaterial Based on Interdigitated and Meander Line Resonator

Rohan Deshmukh

ECE, VNIT

Nagpur, India

rohandeshmukh@students.vnit.ac.in

Dushyant Marathe

Research Scholar, ECE, VNIT

Nagpur, India

dushyantmarathe@gmail.com

Dr. K. D. Kulat

Professor, ECE, VNIT

Nagpur, India

kdkulat@ece.vnit.ac.in

Abstract—We report a new design of multiband electric metamaterial resonator based on integration of interdigitated structure and meander line with square ring. This metamaterial resonator has three distinct electric resonances and negative permittivity regions at C, X band of frequencies. The scattering parameters of proposed sub-wavelength resonator are analysed using full wave electromagnetic simulator Ansys HFSS to demonstrate the presence of electric response at resonant frequencies within 2-12 GHz band. Effective medium parameters permittivity, permeability and refractive index are extracted from simulated scattering parameters. The investigations are also carried out regarding independence of magnetic dipolar activity on flow of surface current. Performance comparison of proposed resonator with single negative SNG ($\epsilon < 0$ and/or $\mu < 0$) resonators is carried out.

I. INTRODUCTION

Metamaterials are the artificially designed and structured media in order to have negative values for permittivity, permeability and refractive index. These negative properties of constitutive parameters (ϵ_r , μ_r) are not found in the materials from which it is made of [1]. The very first realized metamaterial was made from thin wire and split ring resonator (SRR) in order to have overlapping negative permittivity and permeability at bands of microwave frequencies [2]. The metamaterials are classified as ϵ -negative (ENG), μ -negative (MNG) and doubly negative (DNG) based on negative values either for permittivity, permeability or both. This leads to metamaterials with negative refractive index by combining the response from single negative ENG and MNG structures within same unit cell [3]. Applications of single negative metamaterials are found for gain enhancement of antenna [4], design of wave absorber [5], material characterization [6], and microwave sensors [7]. Further, these single negative metamaterials are constituent in the realization of negative refractive index media.

A butterfly pattern resonator [8] shows triple-band response having single and dual bands for negative permittivity and permeability respectively. A triangular shaped single loop resonator [9] has triple band response with two magnetic and one electric resonance. Asymmetric triangular resonator [10] has triple band response with one electric resonance and two magnetic resonances. A compact triple band resonator [11] based on integrating open delta loops with square ring structure has all three distinct electric resonances. A square single loop resonator (square-SLR) [12] is reported to have triple-band response for X-band frequencies. This square-SLR has two distinct magnetic resonances and one very weak electric resonance. Kolb et. al.[13] reported meandered based low resonant frequency structures that has large electric length. The meandered structure resonates at lower frequency and improves effective medium ratio (λ_0/a). The λ_0/a ratio signifies compactness of the metamaterial structures and ascribe effective medium properties to the metamaterials. Kolb has applied successfully the geometrical transformation method on conventional single band E-LC resonator [14] to get low resonant frequency structures. These includes I-shaped, thin square loop, meandered shaped and its variations. The prime disadvantage of the meandered structure is that it has reduced coupling to the external field. A design proposed in this paper is a multiband metamaterial made of square ring with interdigitated and meander line structure having three electric resonances within the frequency range of 6 to 10 GHz. This leads to three well formed negative permittivity regions near resonant frequencies. The transmission and reflection characteristics of proposed resonator are investigated by commercial Finite element method based simulator. In addition, the constitutive effective medium parameters i.e. permeability and permittivity are retrieved from complex scattering parameter. The investigations are carried out regarding flow of time varying surface current at corresponding resonant frequencies to verify strong

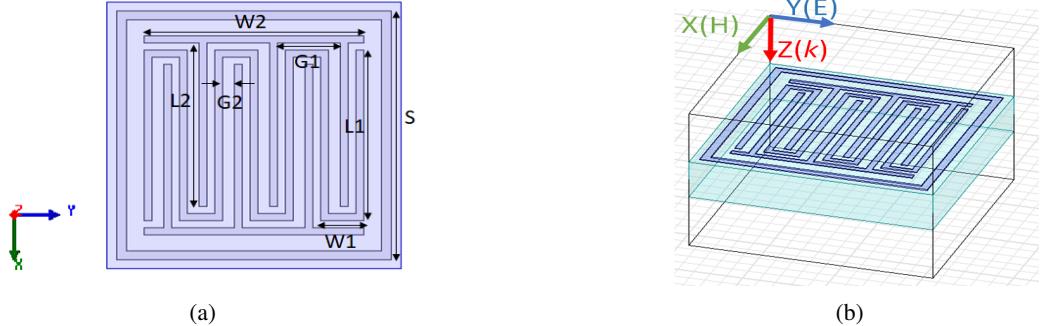


Fig. 1: (a) Geometry of proposed resonator unit cell and (b) Simulation setup for proposed resonator unit cell using PEC-PMC boundaries. For the proposed resonator geometrical parameters are $S = 6.75\text{mm}$, $L_1 = 4.8\text{mm}$, $W_1 = 1.1\text{mm}$, $W_2 = 5.6\text{mm}$, $L_2 = 4.6\text{mm}$, $G_1 = 1.6\text{mm}$, $G_2 = 1.6\text{mm}$.

electric field coupling.

II. DESIGN METHODOLOGY AND SIMULATIONS

The proposed multiband resonator along with its dimensions and the incident electromagnetic wave excitation is shown in Figure 1, where the square ring is combined with interdigitated and meander line. The proposed resonator consists of three different structures namely square ring, interdigitated structure and meander line. The design started with square loop which couples strongly to the electric field to get negative permittivity response for normal wave incidence as shown in [13, 15]. The polarization insensitivity behaviour of square ring for normal wave incidence is indicated in [15]. Apart from this, it is shown that the perpendicular shaped meander line strongly couples to the electric field [13]. Interdigitated structure reported in [16] introduces capacitive effect and provides pure electric resonance.

For square ring, the strip width is 0.25 mm. The width of metal strips of Interdigitated and meander line structure is 0.2 mm. The proposed design is to be printed on single sided Rogers TMM4 dielectric board with copper thickness of 35 μm . The substrate has relative permittivity ϵ_r of 4.7 and thickness of 1.6 mm. The proposed resonator has physical dimensions of $8 \times 8\text{mm}^2$. Copper strips has conductivity of $5.8 \times 10^7 \text{S/m}$.

The electromagnetic response is obtained using finite element method based Ansys HFSS solver version 15. Perfect electric (PEC) and perfect magnetic (PMC) boundary conditions are applied on a unit cell of proposed resonator to perform simulations. Two wave ports model is used to generate source of plane wave excitation along with PEC and PMC boundaries and indicate that unit cells are repeated along X and Y direction. Perfect magnetic and perfect electric boundaries are assigned to the planes parallel to YZ and XZ. This indicates that magnetic field is oriented along X-axis and electric field is oriented along Y-axis as depicted in Figure 1b. We have obtained complex scattering parameters i.e. reflection S_{11}

TABLE I: Simulated resonant frequency f_0 (GHz), transmission minima $S_{21}(\text{dB})$ and effective medium ratio (λ_0/a).

Meta-atom	f_0 (GHz)	S_{21} (dB)	λ_0/a
Proposed resonator	6.18, 7.71 9.36	-46.64, -37.88 -36.24	6.06, 4.86 4

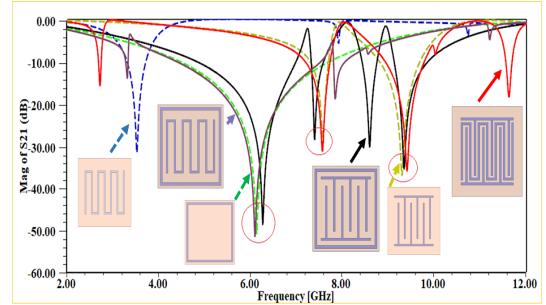


Fig. 2: Transmission spectra of square ring, meander line, interdigitated structure along with mutual structures

and transmission S_{21} from unit cell simulation in the frequency range of 2 to 12 GHz. The transmission spectra of individual resonators i.e. square ring, meander line, interdigitated along with mutual structures is shown in Figure 2 and red circles show the resonance region of proposed resonator. It is observed that the constituent structures i.e. square ring, meander line and interdigitated structure couple strongly with external electric field. The properties of these constituent structures are combined in the proposed resonator in order to get improved electric coupling at all the frequency bands. The magnitude and phase response of the proposed resonator is shown in Figure 3a-3b respectively. The transmission minima $S_{21}(\text{dB})$ is obtained at resonant frequencies 6.18 GHz, 7.71 GHz, and 9.36 GHz as shown in Figure 3a respectively. In Table I, resonant frequencies, S_{21} minima and λ_0/a ratio are listed.

Absorption spectrum of proposed resonator is calculated by

$$A(f) = 1 - |S_{11}(f)|^2 - |S_{21}(f)|^2 \quad (1)$$

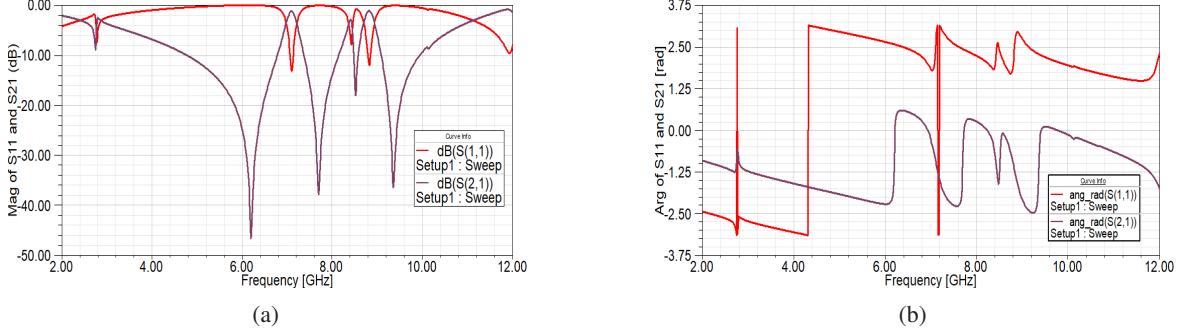


Fig. 3: Simulated transmission and reflection characteristics: (a) magnitude spectra of S_{11} and S_{21} . (b) phase spectra of S_{11} and S_{21} .

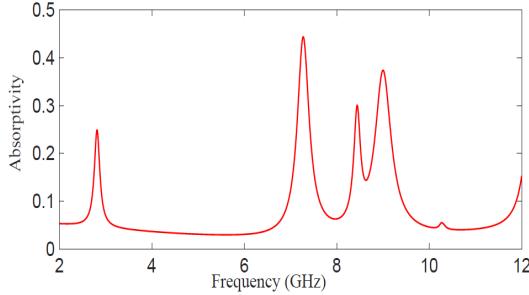


Fig. 4: Absorptivity of proposed resonator

as shown in Figure 4.

Absorptivity is normalized on the scale of 0 to 1. The absorptivity has maximum value of 0.24, 0.44, 0.30 and 0.37 at frequencies 2.81 GHz, 7.27 GHz, 8.44 GHz and 9 GHz respectively for proposed resonator and it can be minimized to some extent by using good conductor metal and high quality substrate.

III. PARAMETER EXTRACTION AND SURFACE CURRENT

The transmission and reflection characteristics of square shaped interdigitated and meander line resonator are obtained from full wave analysis using Ansys HFSS. The parameter extraction method is used to extract complex valued effective medium parameters like effective impedance, refractive index, effective permeability and permittivity. The parameter extraction method is based on [17] and is implemented in Matlab. Parameter extraction methods usually invert Fresnel's relation in order to get effective medium parameters [17, 18, 19]. However, they suffers from disadvantages such as determination of effective thickness of metamaterial slab, branch ambiguity in refractive index due to inversion of complex logarithm. The extraction method reported in [18] consists of direct formulation for refractive index calculation but, it failed to explain the determination of the effective thickness of metamaterial slab. The method reported in [19] has limitation as it employs cumbersome iterative procedure. However, it satisfactorily determines the first effective

boundary and effective thickness of slab. Novelty of parameter extraction algorithm based on Kramers-Kronig relationship [17] is that it computes the real part of refractive index from the imaginary part and overcomes the branch ambiguity issue of the complex logarithm. In this paper, we have employed parameter extraction algorithm based on Kramers-Kronig relationship [17] to determine effective impedance, refractive index and the constitutive parameters as given below

$$z_{eff} = \pm \sqrt{\frac{(1 + S_{11})^2 - S_{21}^2}{(1 - S_{11})^2 - S_{21}^2}} \quad (2)$$

The sign of effective impedance z_{eff} of (2) can be decided by imposing passivity condition given in [20] and threshold value can be adjusted in such a way that if magnitude of z_{eff} is large enough then positive otherwise it is negative. In this case threshold is set to 0.1.

$$e^{in_{eff}k_0d_{eff}} = \frac{S_{21}}{1 - S_{11}R_{01}} \quad (3)$$

where,

$$R_{01} = \frac{z_{eff} - 1}{z_{eff} + 1} \quad (4)$$

Refractive index can be obtained from

$$n_{eff} = \frac{1}{k_0 d_{eff}} \{ Im[\ln(e^{in_{eff}k_0d_{eff}})] + 2m\pi - i Re[\ln(e^{in_{eff}k_0d_{eff}})] \}$$

Where branch index of complex logarithm is denoted by 'm' and it takes value as $m = 0, 1, 2, \dots$. Real part of refractive index can be known from its imaginary part by applying the Kramers-Kronig relation given in [17]

$$n^{KK}(\omega') = 1 + \frac{2}{\pi} P \int_0^\infty \frac{\omega k_{eff}(\omega)}{\omega^2 - \omega'^2} d\omega \quad (5)$$

Where, P denotes the principal value of improper integral. Finally, effective permittivity and effective permeability can be determined from the effective impedance z_{eff} and refractive index n_{eff} as shown below

$$\epsilon_{eff} = \frac{n_{eff}}{z_{eff}} \quad (6)$$

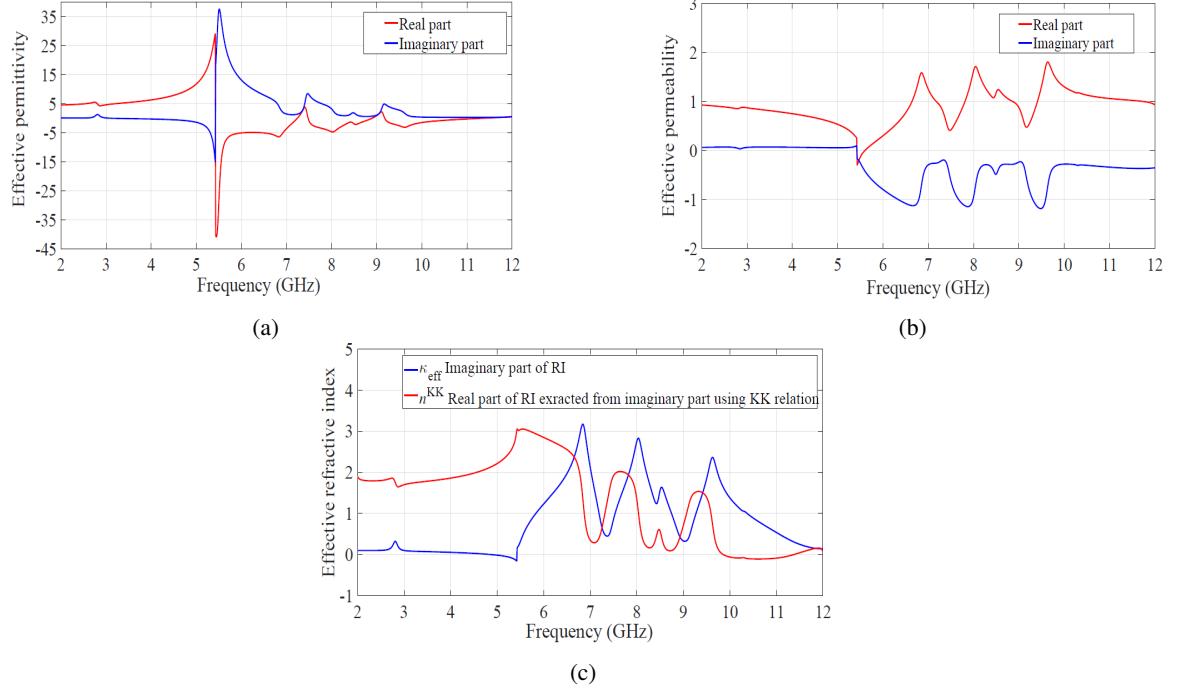


Fig. 5: Computed effective parameters for proposed resonator (a) permittivity, (b) permeability, (c) refractive index. The red line denotes real part and blue line denotes imaginary part.

$$\mu_{eff} = n_{eff} z_{eff} \quad (7)$$

The effective medium parameters are extracted using the above formulation for the proposed resonators as shown in Figure 5a, Figure 5b and Figure 5c respectively. For multiband metamaterial, the electric resonances are clearly seen from the real part of effective permittivity which varies from high positive to negative values near narrow range around resonant frequency as shown in Figure 5a. The real part of effective permittivity has negative values over frequencies 5.4-6.89 GHz, 7.64-8.59 GHz, and 9.19-11.33 GHz for first, second, and third band respectively. The proposed design shows only electric response and no magnetic response as observed from Figure 5a and Figure 5b.

At lower frequency i.e. $f_{01} = 6.18$ GHz, it is observed that the time varying surface current flows through square ring particularly on two horizontal sides as shown in Figure 6a. The surface current flows in clockwise in upper half part of square ring and anti-clockwise in lower half of square ring. This clockwise and anti-clockwise flow of time varying surface ring cancels any magnetic dipolar activity. This gives an idea about resonance at 6.18 GHz of resonator strongly couples with electric field.

At second resonant frequency i.e. $f_{02} = 7.71$ GHz it is observed that the time varying surface current flows through meander line particularly in first two horizontal lines along with the last two horizontal lines. The direction of time varying surface current flow is from left to right as shown in Figure 6b. This flow of surface current

in first two and last two horizontal lines of meander line cancels any magnetic dipolar activity. Same thing follows for vertical lines of meander line, the direction of flow of time varying surface current is in downward direction for second and sixth vertical lines whereas upward direction for fourth and seventh vertical lines.

At third resonant frequency i.e. $f_{03} = 9.36$ GHz it is observed that the time varying surface current flows through interdigitated structure particularly in joints of vertical and horizontal strips of interdigitated structure as shown in Figure 6c. To clearly understand the flow of surface current, divide the interdigitated structure diagonally. For lower triangular part of interdigitated structure flow of time varying surface current is from one vertical branch to other vertical branch through horizontal base line and exactly opposite flow of surface current is observed in upper triangular part of interdigitated structure. This kind of flow of surface current cancels any magnetic dipolar activity and strongly couples with electric field at resonance frequency. All the three different structures strongly couples with electric field and contributes to three different resonances resulted in multiband metamaterial characteristics.

Performance comparison of the proposed resonator with previously reported metamaterial resonators is carried out as shown in Table II. For comparison, we have considered unit cells operating in microwave region. From Table II, it is indicated that, compared to [12, 14, 15, 21], the proposed resonator exhibits triple-band response. At

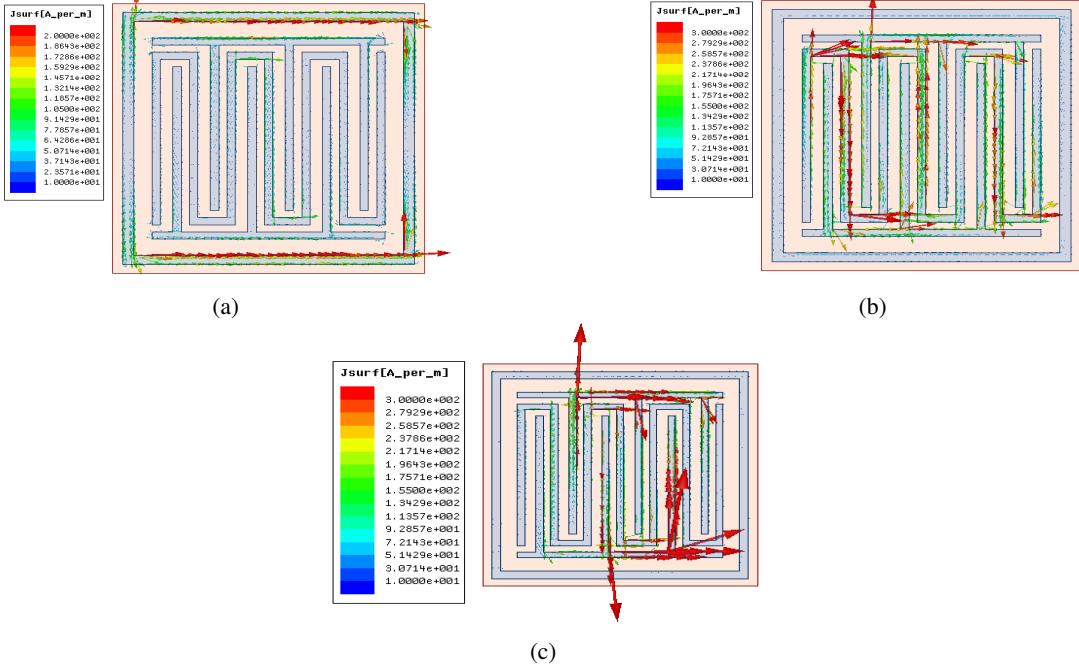


Fig. 6: Surface current at (a) 6.18 GHz (b) 7.71 GHz (c) 9.36 GHz.

TABLE II: Performance comparison with reported single negative SNG ($\epsilon < 0$ and/or $\mu < 0$) metamaterial resonator.

Ref	Geometry	Cell dimensions (mm ²)	f_0 (GHz)	Resonance type	λ_0/a
[14]	E-LC	3.3 × 3.3	15.7	E	5.7
[16]	Idc loaded E-LC	14 × 14	2.11	E	10.2
[21]	Z-shaped	6 × 6	4.9	E	10.2
[22]	Square and meander loaded delta	7.5 × 7.5	5.41, 6.70	E, E	7.39, 5.96
[8]	Butterfly pattern resonator	6 × 8	7.97, 10.31, 12.08	E, M, M	4.7, 3.63, 3.10
[23]	Nested U-ring resonator	6.5 × 6.5	8.24, 9.86, 12.42	M, M, M	5.59, 4.61, 3.70
[24]	Greek-key pattern	10 × 10	2.4, 3.5, 4	E, M, E	12.5, 8.58, 7.45
[11]	Square ring, open delta loops	6 × 6	4.32, 7.55, 9.76	E, E, E	11.45, 6.5, 5.11
Proposed resonator	Square ring with interdigitated and meander line	8 × 8	6.18, 7.71, 9.36	E, E, E	6.06, 4.86, 4

three resonant frequencies, effective medium ratio (λ_0/a) value more than 4 is maintained. The proposed resonator has advantage of having purely electric response as against [8, 23, 24]. Letters E and M in Table II denote type of resonance, namely, electric and magnetic, respectively. Resonator reported in [11] all the three resonances occur due to flow of surface current over top left and bottom right portion of resonator. This results in triple-band response. The proposed resonator resonates at three distinct frequencies with three distinct structures. For lowest resonating frequency square ring resonates, second resonating frequency meander line contributes and third resonance occurs due to interdigitated structure. These are combined together to get three distinct resonances and each resonance is because of the structure which is the main characteristic of the proposed resonator. The fabrication and free space measurements of designed sample will be carried out and included in the future communications.

IV. CONCLUSION

In conclusion, we present a new multiband metamaterial resonator based on square ring with interdigitated and meander line. The scattering parameters and effective medium parameters of proposed metamaterial design are investigated through full wave simulation over frequency range 2 to 12 GHz. Proposed design has purely electric response as expressed by negative permittivity at respective resonant frequencies. The effect of time varying surface current at the resonance is observed and it serves as guideline to get the electric response at resonance. The proposed resonator is expected to find application in realization of multiband devices such as absorbers, filters and antennas over desired frequency band.

REFERENCES

- [1] Viktor G Veselago. The electrodynamics of substances with simultaneously negative values of ϵ and μ . *Soviet physics uspekhi*, 10(4):509, 1968.
- [2] David R Smith, Willie J Padilla, DC Vier, Syrus C Nemat-Nasser, and Seldon Schultz. Compos-

- ite medium with simultaneously negative permeability and permittivity. *Physical review letters*, 84(18):4184, 2000.
- [3] Richard A Shelby, David R Smith, and Seldon Schultz. Experimental verification of a negative index of refraction. *science*, 292(5514):77–79, 2001.
- [4] Dongying Li, Zsolt Szabó, Xianming Qing, Er-Ping Li, and Zhi Ning Chen. A high gain antenna with an optimized metamaterial inspired superstrate. *IEEE transactions on antennas and propagation*, 60(12):6018–6023, 2012.
- [5] Somak Bhattacharyya, Saptarshi Ghosh, and Kumar Vaibhav Srivastava. Bandwidth-enhanced metamaterial absorber using electric field–driven lc resonator for airborne radar applications. *Microwave and Optical Technology Letters*, 55(9):2131–2137, 2013.
- [6] Muhammed Said Boybay and Omar M Ramahi. Erratum to “material characterization using complementary split-ring resonators”[nov 12 3039-3046]. *IEEE Transactions on Instrumentation and Measurement*, 62(6):1866–1866, 2013.
- [7] Jordi Naqui and Ferran Martín. Angular displacement and velocity sensors based on electric-lc (elc) loaded microstrip lines. *IEEE Sensors Journal*, 14(4):939–940, 2014.
- [8] Guohong Du and Changjun Liu. Multiband metamaterial structure: Butterfly-pattern resonator. *Microwave and Optical Technology Letters*, 54(9):2179–2181, 2012.
- [9] Ozan Yurduseven, Asim Egemen Yilmaz, and Gonul Turhan-Sayan. Triangular-shaped single-loop resonator: a triple-band metamaterial with mng and eng regions in s/c bands. *IEEE Antennas and Wireless Propagation Letters*, 10:701–704, 2011.
- [10] Cheng Zhu, Jing-Jing Ma, Long Li, and Chang-Hong Liang. Multiresonant metamaterial based on asymmetric triangular electromagnetic resonators. *IEEE Antennas and Wireless Propagation Letters*, 9:99–102, 2010.
- [11] Dushyant Marathe and Kishore Kulat. A compact triple-band negative permittivity metamaterial for c, x-band applications. *International Journal of Antennas and Propagation*, 2017, 2017.
- [12] E Ekmekci and G Turhan-Sayan. Single loop resonator: dual-band magnetic metamaterial structure. *Electronics letters*, 46(5):324–325, 2010.
- [13] PW Kolb, TS Salter, JA McGee, HD Drew, and WJ Padilla. Extreme subwavelength electric ghz metamaterials. *Journal of Applied Physics*, 110(5):054906, 2011.
- [14] D Schurig, JJ Mock, and DR Smith. Electric-field-coupled resonators for negative permittivity metamaterials. *Applied physics letters*, 88(4):041109, 2006.
- [15] Biao Li, Lianxing He, Yingzeng Yin, Wanyi Guo, and Xiaowei Sun. An isotropy dual-band terahertz metamaterial. *Microwave and Optical Technology Letters*, 55(5):988–990, 2013.
- [16] Withawat Withayachumnankul, Christophe Fumeaux, and Derek Abbott. Compact electric-lc resonators for metamaterials. *Optics Express*, 18(25):25912–25921, 2010.
- [17] Zsolt Szabó, Gi-Ho Park, Ravi Hedge, and Er-Ping Li. A unique extraction of metamaterial parameters based on kramers-kronig relationship. *IEEE Transactions on Microwave Theory and Techniques*, 58(10):2646–2653, 2010.
- [18] DR Smith, DC Vier, Th Koschny, and CM Soukoulis. Electromagnetic parameter retrieval from inhomogeneous metamaterials. *Physical review E*, 71(3):036617, 2005.
- [19] Xudong Chen, Tomasz M Grzegorczyk, Bae-Ian Wu, Joe Pacheco Jr, and Jin Au Kong. Robust method to retrieve the constitutive effective parameters of metamaterials. *Physical review E*, 70(1):016608, 2004.
- [20] Hongsheng Chen, Lixin Ran, Jiangtao Huangfu, Xianmin Zhang, Kangsheng Chen, Tomasz M Grzegorczyk, and Jin Au Kong. Left-handed materials composed of only s-shaped resonators. *Physical Review E*, 70(5):057605, 2004.
- [21] Abdallah Dhouibi, Shah Nawaz Burokur, André de Lustrac, and Alain Priou. Z-shaped meta-atom for negative permittivity metamaterials. *Applied Physics A*, 106(1):47–51, 2012.
- [22] Abhishek Sarkhel, Debasis Mitra, Sandip Paul, and Sekhar Ranjan Bhadra Chaudhuri. A compact meta-atom for dual band negative permittivity metamaterial. *Microwave and Optical Technology Letters*, 57(5):1152–1156, 2015.
- [23] O Turkmen, E Ekmekci, and G Turhan-Sayan. Nested u-ring resonators: a novel multi-band metamaterial design in microwave region. *IET microwaves, antennas & propagation*, 6(10):1102–1108, 2012.
- [24] Behnam Zarghooni, Abdolmehdi Dadgarpour, and Tayeb A Denidni. Greek-key pattern as a miniaturized multiband metamaterial unit-cell. *IEEE Antennas and Wirel. Propag. Lett*, 14:1254–1257, 2015.

Design of Frequency-Signature Based Multiresonators Using Quarter Wavelength Open Ended Stub for Chipless RFID Tag

P Prabavathi¹, S Subha Rani²,

¹Assistant Professor, ²Professor & Senior Member IEEE,

Department of Electronics and Communication Engineering, PSG College of Technology, Coimbatore, Tamilnadu, India
pprabavathi@gmail.com, ssr.ece@psgtech.ac.in

Abstract— A quarter wavelength open stub multiresonators are proposed for chipless Radio Frequency Identification (RFID) tag. The data capacity of the tag is 10 bit and open stub resonator operates in the frequency band of 2 GHz to 4 GHz. The data in the tag is encoded using absence or presence coding and frequency shift coding (FSC). The data stored is read and transmitted using the planar circular patch monopole ultra-wide band (UWB) antenna. The tag consists of two cross polarized sending and receiving planar circular patch (PCP) monopole UWB antennas connected to the multiresonators. The span of the multiresonator based chipless RFID tag is 23.8mmx17mm which is designed on FR4 substrate with dielectric permittivity of 4.4 and tangent loss of 0.01. It is designed and tested under simulation using ADS software and the vector network analyzer (VNA) after fabrication. The tag has the insertion loss in the range of -10 dB to -30 dB and a bit density of 2.47bits/cm².

Keywords—Chipless Radio Frequency Identification (RFID) tag, coding techniques, open stub multiresonator, PCP monopole UWB antenna.

I. INTRODUCTION

The radio frequency identification is the future generation of bar code technology used for electronic product identification. The tag is used to encode the data of the product used in variety of fields. RFID technology has proven advantages than the existing bar code identification system. They are detected many at a time through obstacles. They are also more robust and reliable than the former system. There are two types of RFID tags, active and passive. Passive tag does not contain internal power supply and the operating range is small. Active tags come with an inbuilt power supply and it can be read over some kilometers distance.

Usually, the RFID tag consists of an ASIC silicon chip in which the data is stored. The chipless RFID tag can be designed without the need for a silicon chip for data storage. Chipless RFID tag instead makes use of the multiresonators for

efficiently storing and retrieving the data. The RFID system includes RFID tag, middleware software and a reader.

The RFID tag can be either operated in time domain or in frequency domain. In [1], the time domain reflectometry (TDR) based tags, reflections of the signal from the piezoelectric substrate or a crystal are used for encoding the data and spectral signature-based chipless RFID tags, frequency domain is used for encoding the data. They use particular frequency bands for data encryption. They can be either multiresonator based tags or multiscatterer based tags. The multiresonator based tag consists of a multiresonator circuit, a vertically polarized antenna and a horizontally polarized antenna [2]. The multiscatterer tag does not need an antenna as the different sized resonating structures are used for transmission and reception of the data. RFID reader is used to read the data stored in the tag done by receiving the RF signals. The middleware software is used for decoding the data from the reader to the personal computer.

The multiresonators can be designed by selecting different sizes and shapes of resonators. Multiresonators are structures which resonate at different predetermined frequencies over its operating range. The data bit can be stored in its resonant peaks through absence or presence coding technique. The data can be read by observing the frequencies in which they resonate in through frequency shift coding technique.

In [3], a complete chipless RFID tag using shorted stub resonators are designed. The L shaped microstrip open stub multiresonators are designed [4]. An open stub band stop filter with spurline is designed [5]. A dual-band band stop filter is designed with two embedded open stubs and one spurline [6]. Circular shaped nested slots are proposed in [7]. The different shapes of multiple microstrip resonator are used for spectral signature is proposed in [8]-[12].

The integration of dual-band resonators in frequency coded tags is proposed in [13]. Spurline

resonators having stubs with quarter wavelength are used in [14], [15]. Multiple spurline based band stop filter is designed [16].

II. MULTIRESONATOR DESIGN

The chipless RFID tag can be designed with the help of multiresonators. In [3], a shorted stub multiple resonators are designed for chipless RFID tag. A half wavelength stub is shorted to ground through a hole by connected its far end to the microstrip transmission line. The bottom plane acts as ground.

The shorted stub resonator is half wavelength long at the predetermined resonant frequency. A shorted stub is less affected by the leakage of electromagnetic radiation. The gain is stable over wide range of frequency. The multiresonator size is large. It is difficult to fabricate since it uses via hole connection.

To overcome the limitations of the shorted stub multiresonators, various open stub multiresonators are proposed in this paper. They achieve higher gain with reduced size. The simple open stub multiresonator with quarter wavelength is designed and taken as the basic structure. Hence, 50% of the tag size can be reduced on designing open stubs multi resonators.

A. Simple Open Stub Multiresonator

The multiresonator consists of a 50 ohm impedance microstrip transmission line to which the stubs are attached. Each stub can encode a bit. Since the data encoding capacity of the tag is 10 bits, maximum of 10 stubs are used. The tags are encoded using absence or presence coding technique. Each open stub is designed such that it resonates at a predetermined frequency.

When the stub is present, it resonates at that designed frequency and hence the data bit is encoded as 1. When the stub is absent, there is no resonance at that frequency which indicates that the data bit is encoded as 0. Since the coding is done in the frequency domain, it is included under spectral signature based chipless RFID tags.

A quarter wavelength open stub is connected with its far end to the microstrip transmission line. The bottom plane acts as ground.

The open stub microstrip line resonator is quarter wavelength long at the resonant frequency.

$$W_i = \frac{\lambda_g}{4} \quad (1)$$

where λ_g is the guide wavelength, W_i is stub length,

$$\lambda_g \approx \frac{\lambda}{\sqrt{\epsilon_{eff}}} \text{ where } \epsilon_{eff} \text{ is the effective permittivity}$$

of the substrate.

The span of a 10 bit multiresonator structure is 23.8mmx17mm as shown in Fig. 1. For the resonant frequency of 2.06 GHz, the open stub length is about 20.8 mm. The same concept is used for all the stubs in the multiresonators.

The FR4 substrate is used for layout design with the dielectric permittivity of 4.4 and tangent loss of 0.01. The substrate height is taken as 1.6 mm.

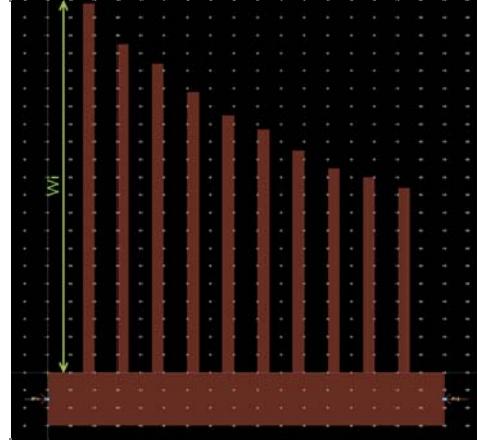


Fig.1 Layout of a 10 bit simple open stub multiresonator

The conductor thickness is about 30 μm . The length of the transmission line is 17 mm and width is about 3 mm. The open stub resonators mutual coupling are reduced by stubs are kept 1mm apart.

B. Multiresonator with L shaped slot in the transmission line

To further reduce the amount of conductor to be used in the multiresonator, slots can be cut in the microstrip transmission line. Care should be taken such that the gain is maintained nearly same as the simple open stub multiresonator.

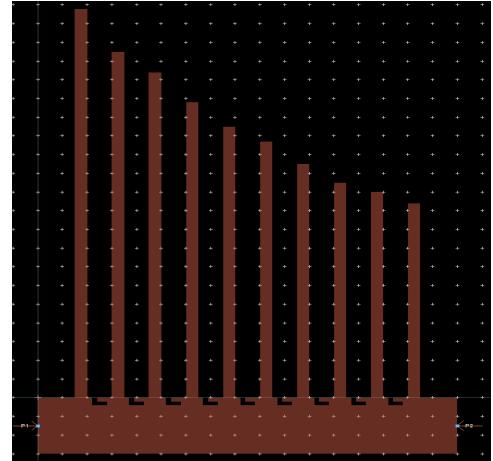


Fig.2. Layout of a 10 bit open stub multiresonator with slot in the transmission line

The ‘L’ shaped slot has the length of 0.4 mm in its vertical arm and 0.6 mm in its horizontal arm and width of the slot is 0.2 mm. The ‘L’ shaped slots are made between each pair of stubs attached in the microstrip transmission line. Hence there are nine slots in the multiresonator in total. The amount of the conductor is reduced and also no surface current will be induced around the slot with some compensation in insertion gain.

C. Length of the stubs used in multiresonator design

The length of the stubs used in the various open stub multiresonators are calculated using the equations (1). It is also noted that the length of the stubs used for slotted techniques are also the same.

The lengths are tabulated in the Table I.

TABLE I. LENGTH OF THE STUBS

I	Resonant frequency (in GHz)	Length of the open stub W_i (in mm)
1	2.08	20.8
2	2.22	18.5
3	2.36	17.4
4	2.59	15.8
5	2.80	14.5
6	2.96	13.7
7	3.19	12.5
8	3.41	11.5
9	3.68	11
10	3.84	10.4

III. FREQUENCY SHIFT CODING TECHNIQUE

In this coding technique, several data bits is used to encode the resonance using a single stub. This is far more effective than the absence or presence coding technique. It is based on the principle that, by increasing the length of each stub, it can be made to resonate at different frequency. The increase in length has to be smaller such that the resonant frequencies are packed in a finite bandwidth.

More number of frequency values can be represented in a small definite bandwidth if their change in length is small. The frequency shift has to be as small as possible. Fig.3. shows the structure of open stub multiresonator used for the frequency shift coding.

In this open stub multiresonator, four stubs are being used for FSC. Each stub can represent 9 different data symbols combination. The data representation is taken as from 0 to 8. Hence, each stub is used for representation of anyone out of nine data symbols. In this technique, 6561 different identification codes are generated with different data bit combinations.

In this technique, four stubs are used in the representation of data bits. Each stub can represent a

bit. A frequency band of 2.03 GHz to 2.20 GHz is allotted to first stub, 2.39 GHz to 2.66 GHz is allotted to second stub, 2.74 GHz to 3.11 GHz is allotted to third stub and 3.24 GHz to 3.70 GHz is allotted to fourth stub.

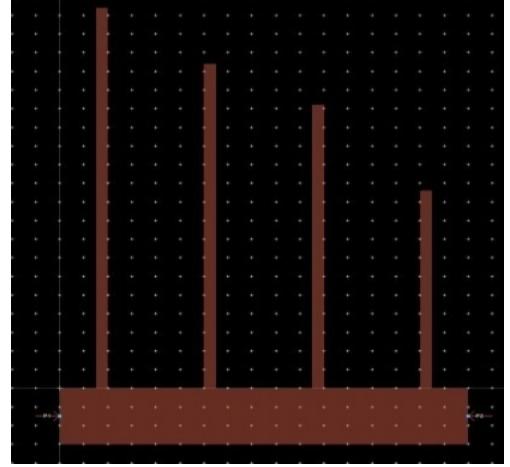


Fig.3. Layout of open stub multiresonator used for frequency shift coding

IV. ANTENNA DESIGN

The RFID tag consists of the antenna for transmission and reception of data from the multiresonator [2]. The multiresonator circuit is connected to the cross polarizing PCP monopole UWB antenna. Any antenna which operates in the frequency range of 3.1 GHz to 10.6 GHz is termed as UWB antenna. One such antenna designed for the multiresonator operation is PCP monopole UWB antenna. So here two antennas are needed for a complete chipless RFID tag, One for reception of the signal another for transmission the signal.

The antenna should have good impedance which matches the operating frequency of the multiresonator. Due to simple structure and wideband operation, the PCP UWB monopole antenna can be selected.

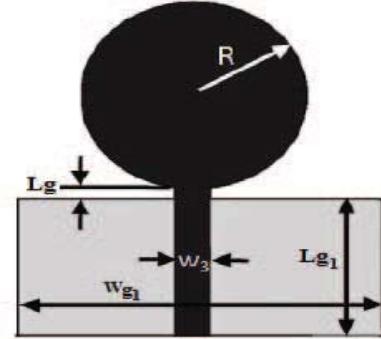


Fig.4. Geometry of the UWB disc monopole antenna

UWB systems have larger bandwidth. The signal is robust and the implementation is simple. Fading can

be reduced in this band. They utilize less power and inexpensive. Hence, an UWB PCP monopole antenna is selected for the purpose.

The geometry of the UWB PCP monopole antenna is shown in Fig. 4. A PCP monopole antenna consists of a circular patch of radius ‘ r ’ fed with microstrip line printed on dielectric material. Omni-directional radiation pattern allows exciting even from the back side of the tags. W_3 is the width of the feed line (microstrip). W_{g1} and L_{g1} denote width and length of the partial ground plane (rectangular structure) respectively. The partial ground plane and circular patch gap is L_g .

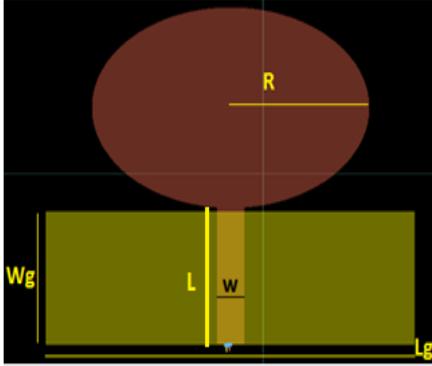


Fig.5. Layout of the UWB disc monopole antenna ($R=15$ mm, $W=3$ mm, $L_g=40$ mm, $W_g=25$ mm, $L=15$ mm, $\epsilon_r=4.4$, height of the substrate= 1.6 mm)

Fig. 5. shows the layout of the UWB PCP monopole antenna. The final structure of the tag consists of a vertically polarized PCP UWB monopole receiving antenna, a multiresonator with 10 stubs and a horizontally polarized PCP UWB monopole transmitting antenna. The two cross polarized monopole UWB antennas are used to reduce the interference between the signals (transmitted and received).

V. RESULTS AND DISCUSSIONS

The simulated results for open stub multiresonators and the PCP monopole antenna are discussed in this section. The multiresonator is simulated using the ADS software. The simulation parameters are chosen as mentioned above.

A. Simple Open Stub Multiresonator

The fabricated multiresonator on FR4 substrate with the loss tangent of 0.01 and dielectric permittivity of 4.4 is shown in the Fig. 6. The insertion loss characteristic shows the resonance at different predetermined frequencies for various stubs in the multiresonator. The simulated and measured response of the multiresonator for the bit combination 1111111111 is shown in Fig. 7. Hence this multiresonator circuit design generates unique identification code 1111111111. The product detail

which matches with the respective code can be detected using the RFID reader.

The multiresonator is connected to VNA through the SMA connector. The insertion loss (S_{21}) characteristics taken in VNA are compared with the simulated results in ADS as shown in the Fig. 7. The multiresonator measured results observed that there is a small frequency shift due to fabrication.

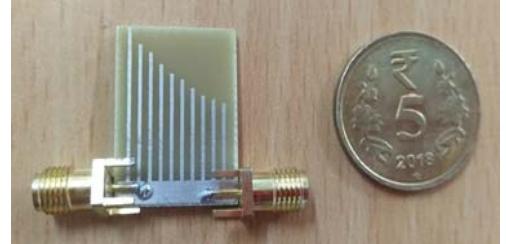


Fig.6. Fabricated open stub multiresonator

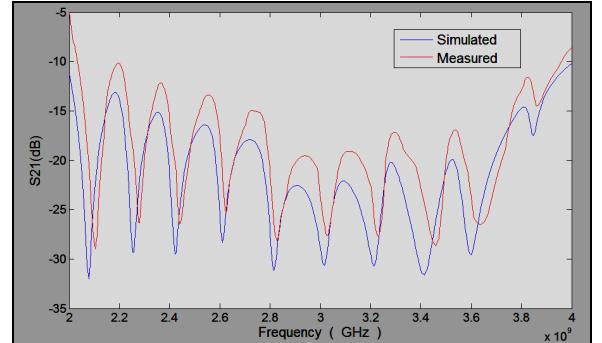


Fig.7. S_{21} magnitude of a 10 bit simple open stub multiresonator

The resonant frequencies of the multiresonator are 2.08 GHz, 2.22 GHz, 2.36 GHz, 2.59 GHz, 2.80GHz, 2.96 GHz, 3.19 GHz, 3.41 GHz, 3.68 GHz, 3.84 GHz.

The modified structure of multiresonator circuit design contains ten microstrip open stub resonators. In the absence or presence of resonance coding technique, the 10-bit multiresonator circuit design generated 1024 unique identification codes.

The tag has the insertion loss in the range of -10 dB to -30 dB except at the last frequency. This is due to the instability of the open stubs at higher frequencies.

B. Multiresonator with L shaped slot in the transmission line

The fabricated multiresonator shown in Fig. 8 is tested using VNA and results are compared with that of the simulation as shown in Fig. 9. The simulated insertion loss characteristic shows the resonance at different frequencies for multiresonator with slots in the microstrip transmission line. The simulated and measured response of the multiresonator for the data bit combination 1111111111 is shown in Fig. 9 which

is the unique identification code of the multiresonator circuit.

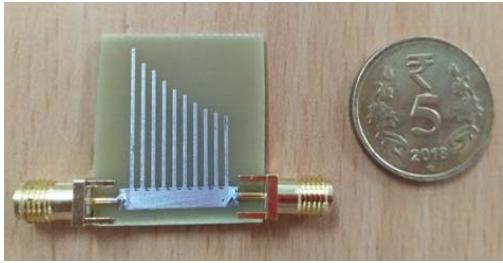


Fig.8. Fabricated open stub multiresonator with slots in the transmission line

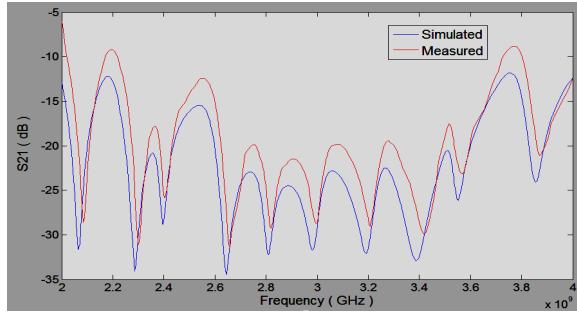


Fig.9. S_{21} magnitude of a 10 bit open stub multiresonator with slots in the transmission line

The result shows that there is small resonance frequency shift due to change in capacitance. However, the gain is higher than the other previously used techniques.

C. Frequency Shift Coding Technique

Initially, the length of the first stub is only varied while length of the other stubs remains constant. Each of the stubs has resonant frequency that can vary between any nine values. The first stub length is taken as 20.6 mm and the length is kept increasing by 2 mm to produce nine different resonant frequencies.

The same procedure is repeated for all the other three stubs. The results are shown for the tuning of first stub in the Fig. 10, for the tuning of second stub in the Fig. 11, for the tuning of third stub in the Fig. 12, and for the tuning of fourth stub in the Fig.13.

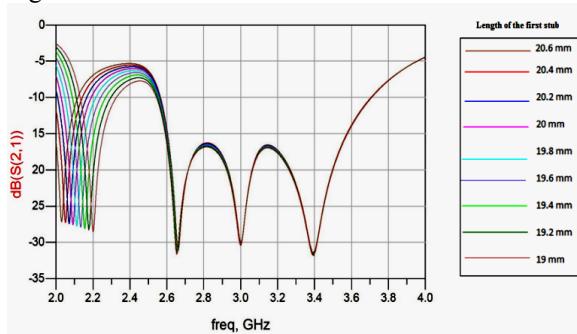


Fig.10. Simulated S_{21} magnitude of tuning of first stub

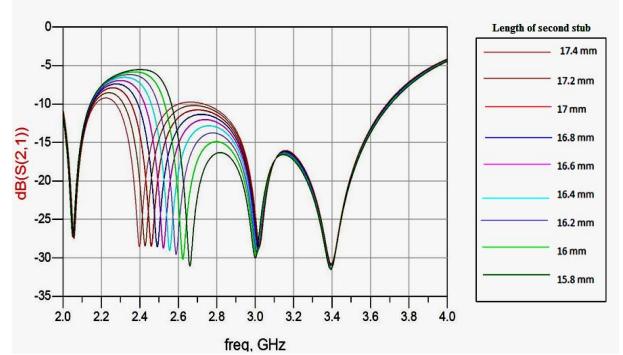


Fig.11. Simulated S_{21} magnitude of tuning of second stub

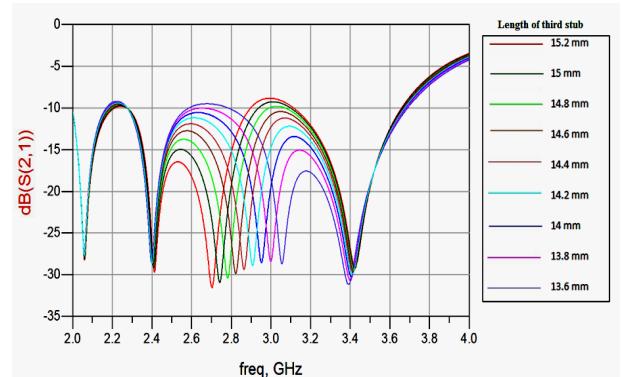


Fig.12. Simulated S_{21} magnitude of tuning of third stub

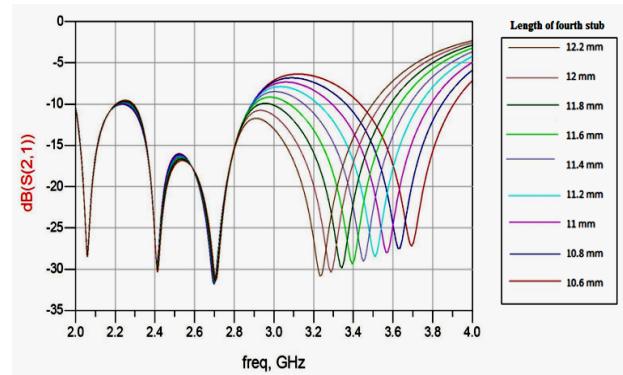


Fig.13. Simulated S_{21} magnitude of tuning of fourth stub

D. Antenna measurement

A reflection loss characteristic of the antenna shows that very good performance. The designed multiresonator is operated in the frequency range of 2GHz to 4GHz. Hence, the antenna must exhibit a good performance over this range.

The PCP UWB antenna is made to resonate at several frequencies over the range of 2GHz to 6GHz. The antenna should have the S_{11} magnitude to be at minimum of -10 dB to perform well in the frequency range. The fabricated UWB monopole antenna is shown in Fig. 14. Measured reflection loss characteristics by VNA are shown in Fig. 15.



Fig.14. Fabricated disc monopole antenna

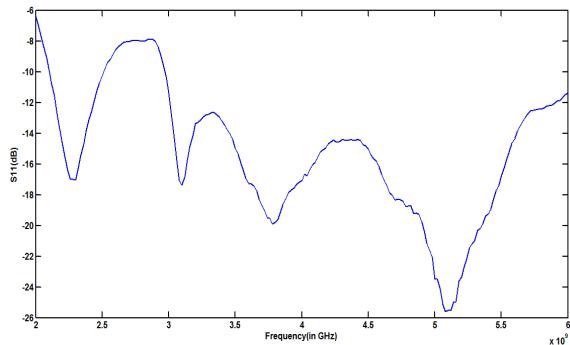


Fig.15. Measured S_{11} magnitude of the monopole antenna

VI. CONCLUSION

In this paper, quarter wavelength open stub multiresonators and multiresonators with 'L' slots in the microstrip transmission line are proposed. The tradeoff existing between the reduced metal requirement and the insertion gain are discussed for each multiresonator. They can be chosen appropriately according to our specific needs. However, all the designed tag is operated in the frequency band of 2 GHz to 4 GHz which makes them flexible to use for various applications. In FSC technique 6561 identification codes are generated in comparison with the absence or presence coding technique which generates 1024 identification codes. The antenna operating over the range of 2 GHz to 6 GHz also exhibits good performance.

REFERENCES

- [1] Preradovic S, Karmakar N. C, "Chipless RFID: Bar Code of the Future", IEEE Microwave Magazine, Dec 2010, p. 87-96.
- [2] S. Preradovic, I. Balbin, N. C. Karmakar, and G. Swiegers, "A novel chipless RFID system based on planar multiresonators for barcode replacement," in Proc. IEEE Int. Conf. RFID, Las Vegas, Apr. 16–17, 2008, pp. 289–296.
- [3] M Sumi, R Dinesh, C M Nijas, P Mohanan, S Mridula, "Frequency signature based Chipless RFID tag using shorted stub resonators," IEEE 4th Asia-Pacific Conference on Antennas and Propagation (APCAP), July 2015.
- [4] Nijas, C. M., R. Dinesh, U. Deepak, Abdul Rasheed, S. Mridula, K. Vasudevan, and P. Mohanan. "Chipless RFID tag using multiple microstrip open stub resonators," *IEEE Transactions on Antennas and Propagation*, Volume 60, no. 9, Sep. 2012, pp. 4429-4432.
- [5] Tu, Wen-Hua, and Kai Chang. "Compact microstrip bandstop filter using open stub and spurline." *IEEE Microwave and Wireless Components Letters*, Volume 15, no. 4, Apr 2005, pp. 268-270.
- [6] Yang, S., "A compact dual-band bandstop filter having one spurline and two embedded open stubs," *Journal of Electrical Systems and Information Technology*, Volume 3, Issue 2, pp. 314-319, Sep. 2016.
- [7] Martinez, Marcos, and Daniel van der Weide, "Compact slot-based chipless RFID tag," *IEEE RFID Technology and Applications Conference (RFID-TA)*, pp. 233-236, Sep. 2014.
- [8] Martinez-Iranzo, Ursula, Bahareh Moradi, and Joan Garcia-Garcia, "Open ring resonator structure for compact chipless RFID tags," In *Microwave Symposium (IMS), IEEE MTT-S International*, pp. 1-3, Jul. 2015.
- [9] Mohanan, Dinesh Rand P. "Spectral Signature based Chipless RFID Tag using Coupled Bunch Resonators," *European Journal of Advances in Engineering and Technology*, Volume 2, no. 11, 2015
- [10] Sumi, Mohan, Raghavan Dinesh, Chakkanattu Mohammedkunju Nijas, Shanta Mridula, and Pezhohil Mohanan, "High Bit Encoding Chipless RFID Tag Using Multiple E Shaped Microstrip Resonators," *Progress In Electromagnetics Research B* 61 (2014): 185-196.
- [11] Dinesh, R., P. V. Anila, C. M. Nijas, M. Sumi, and P. Mohanan, "Open loop multi-resonator based chipless RFID tag," *General Assembly and Scientific Symposium (URSI GASS), 2014 XXXIth URSI*, pp. 1-4. IEEE, 2014.
- [12] Gu, Q., G. C. Wan, C. Gao, and M. S. Tong, "Frequency-coded chipless RFID tag based on spiral resonators," *Progress in Electromagnetic Research Symposium (PIERS)*, pp. 844-846. IEEE, 2016.
- [13] Girbau, David, Javier Lorenzo, Antonio Lazaro, Carles Ferrater, and Ramón Villarino. "Frequency-coded chipless RFID tag based on dual-band resonators." *IEEE Antennas and Wireless Propagation Letters*, Volume 11, Jan. 2012, pp. 126-128.
- [14] Sumi, M., R. Dinesh, C. M. Nijas, S. Mridula, and P. Mohanan, "Frequency coded chipless RFID tag using spurline resonators," *Radio Eng* 24, no. 4 (2014): 203-208.
- [15] Angkawisitpan, Niwat, "Miniaturization of bandstop filter using double spurlines and double stubs," *Przeglad Elektrotechniczny* 88, no. 11a (2012): 178-181.
- [16] Shrestha, Bhanu, and Surendra Shrestha, "Miniaturized Multi-spurline Bandstop Filter Design with a Meandered Slot Lines," *Journal of the Institute of Engineering* 11, no. 1 (2016): 172-176.

Slot Antenna Miniaturization Using Copper Coated Circular Dielectric Material

Enamul Khan¹, Jinia Aktar², Khan Masood Parvez³ and SK. Moinul Haque⁴

Department of Electronics and Communication Engineering

Aliah University

Newtown, Kolkata, West Bengal-700160, India

¹enamulrph85@gmail.com,²jinia13417@gmail.com,³masoodrph@gmail.com,⁴moinul3@gmail.com

Abstract— The contribution of this paper is to proposed a simple slot antenna miniaturization method using copper coated FR_4 dielectric material loading technique. The operating frequency for reference antenna and loaded antenna are 2.86GHz and 1.66GHz respectively. As a result, overall size of proposed antenna reduces by a ratio of 41.95%. A parametric study on various copper coated dielectric materials is presented to better understand the effect of the permittivity on slot antenna miniaturization. The antenna topologies are designed and analyzed using High Frequency Structure simulator (HFSS) tool. The prototype was fabricated and measured, and the measured results show good agreements with the simulated ones. This antenna can be very useful for various wireless communication systems.

Keywords— slot, copper coated, dielectric material, microstrip fed, miniaturized.

I. INTRODUCTION

In recent times, there is the huge development in antenna miniaturization techniques for past few decades due to its wide range of application in electronics gadgets, advanced defense and security system. The electronic capsules are also required miniaturized antenna for monitoring and diagnosis the human body. Satellite industries are also moving towards the small and nano-satellite to reduce the cost of the system. Antenna radiation features are strongly linked with physical size which is comparable to its wavelength. Then reduction of physical size or operating frequency of an antenna keeping the other parameter unchanged is not an easy task for scientific community because it hampers antenna's return loss, gain, radiation pattern, efficiency. However, there are several techniques to achieve miniaturization using high permittivity, high permeability dielectric material, artificial magnetic conductor, optimize the antenna geometry or shapes.

The slits, strips and loops loading techniques for slot antenna miniaturization have been addressed in literature [1]. The inductive loading offered by slits, strips and loops causes 48.01% reduction in resonant frequency. The antenna efficiency is also achieved more than 90% along with good radiation pattern. Several microstrip patch antenna miniaturization methods were also reported in [2]. These methods are the metal loading, shorting and folding, introducing slot, modified ground plane and metamaterials. Recently, the 3-D printed quadrifilar helix antenna miniaturization technique [3] reduces the antenna size of a ratio of 43%. The effect of slot antenna miniaturization using circular, triangular and rectangular loops with equal

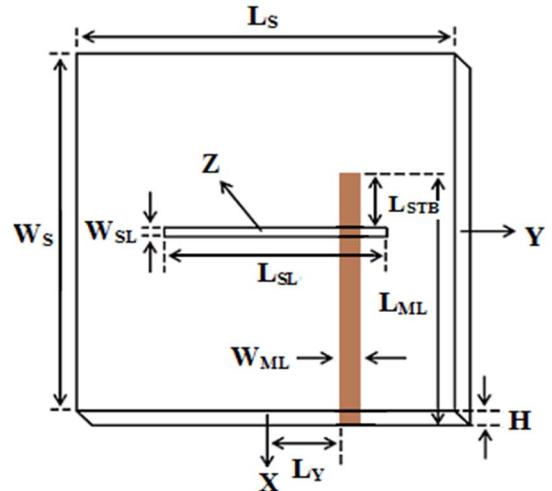


Fig.1. Reference antenna.

electrical path length has been described in [4]. Another unique way of CPW fed slot antenna miniaturization is denoted in [5]. The resonant frequency reduces by a factor of 37.73% for slits on both sides of the slot and 42.33% for slits on fed side of the slot. In [6], an analytical expression is presented for the quality factor of an antenna loaded with a combination of two different materials. The loop loading miniaturization techniques for dipole, monopole and slot antenna have been addressed in [7]. In addition, the MNG metamaterial [8] is used to reduce the size of circular patch antenna. The design of loop and dielectric resonator loaded antenna topology is presented in [9]. Novel approach in slot antenna miniaturization and enhanced bandwidth have been noted in [10], [11].

This paper highlights a new way of slot antenna miniaturization using copper coated FR_4 dielectric material loading technique. The slot dimensions and dielectric substrate properties are kept same for reference antenna and loaded antenna. The copper coated dielectric material create the inductive situation in reference antenna topology which is cancelled out in capacitance reactance present at reference antenna resonant frequency causing a huge reduction in resonant frequency verified by simulated impedance verses frequency plot.

II. ANTENNA DESIGN

The microstrip fed slot antenna is presented in Fig.1 as a reference antenna to better comparison with copper coated dielectric loading effect in antenna miniaturization. The

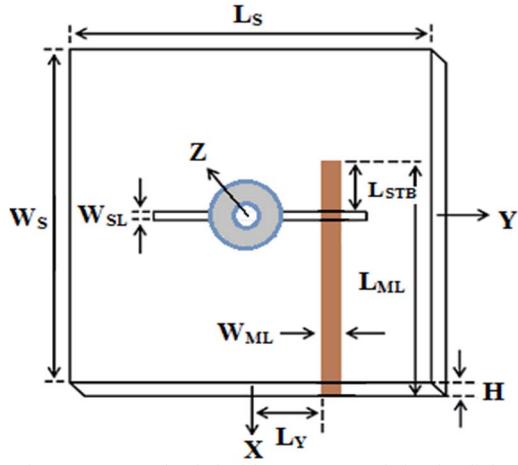


Fig.2. Reference antenna loaded with copper coated circular dielectric material

substrate used for designing this antenna configuration is flame retardant (FR_4) has the permittivity and loss tangent of 4.4 and 0.02 respectively. The antenna topology has the following substrate dimensions: substrate length (L_s) = substrate width (W_s) = 100.00 mm, substrate height (H) = 1.58mm. A slot antenna consists of a conductive ground plane slot cut out at top of the dielectric substrate. The slot having the length (L_{SL}) of 34.00mm and Width (W_{SL}) of 1.20mm is used to radiate the electromagnetic energy at 2.86 GHz due to discontinuity of current in conductive surface. The microstrip line is placed on bottom of the substrate to excite the slot. The length (L_{ML}) and trace width (W_{ML}) of microstrip line are 52.20mm and 2.95mm respectively. The stub length (L_{STB}) and distance (L_y) from origin in y direction are 1.6 mm and 7.85 mm respectively. In general, microstrip fed slot antenna is most broadly used for wireless communication system due to its smart structure like compactness, low-profile, less expensive and easy to implementation with others devices. The antenna topologies are simulated using High Frequency Structure Simulator (HFSS) tool for numerical solution [12].

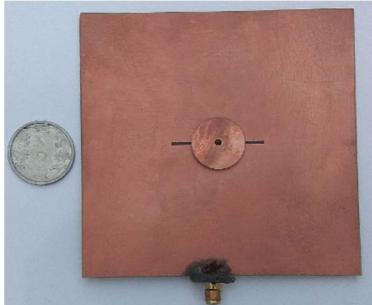


Fig.3. Fabricated prototype of reference antenna loaded with copper coated circular dielectric material

The reference antenna (Fig.1) is loaded with copper coated circular FR_4 dielectric material at the center of the slot with a hole. The top surface of circular dielectric material is coated with copper. The antenna configuration is depicted in Fig.2. The thickness of FR_4 material is 1.58mm. The circular dielectric material has diameter (D_1) of 20 mm. The hole is filling up with air. The diameter (D_2) of hole is 2mm. The microstrip line length (L_{ML}), microstrip trace width (W_{ML}) are taken 54.20 mm and 2.95 mm respectively.

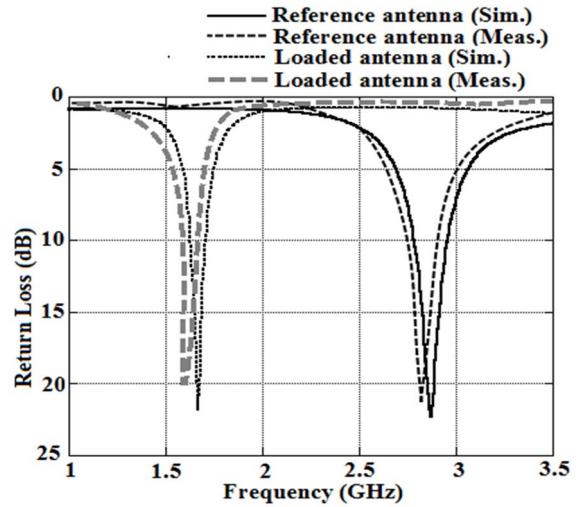


Fig. 4. Simulated and measured return loss of loaded and reference antenna topology

Also, the microstrip line stub length (L_{STB}) and distance (L_y) from the origin in Y direction are 3.60 mm and 1.50 mm respectively. The fabricated prototype is shown in Fig.3 to understand the proper antenna configuration.

III. RESULTS AND DISCUSSIONS

The simulated and measured return loss characteristics of loaded antenna are shown in Fig.4. The simulated values of resonant frequencies for loaded and reference antennas are 1.66 GHz and 2.86 GHz respectively where as measured values are 1.61 GHz and 2.82 GHz respectively. It can be found that simulated and measured values are agreeing well with each other. The percentage of reduction in resonant frequency is 41.95%. Also, the -10 dB bandwidth for loaded and reference antennas are 4.20% and 5.93% respectively which produces 29.17% degradation in -10 dB bandwidth

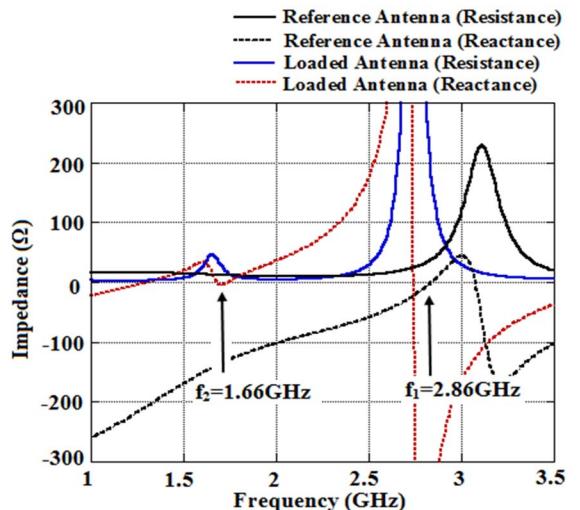
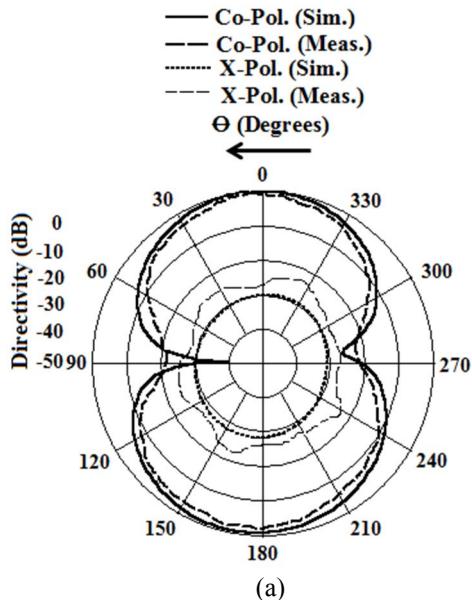
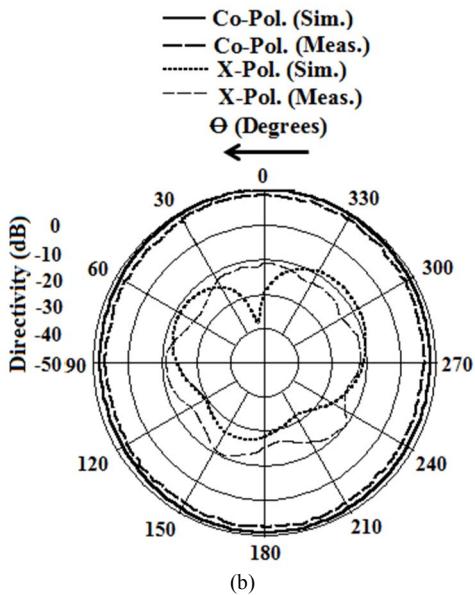


Fig.5. Simulated input impedance verses frequency plot for reference antenna and loaded antenna

The input impedance characteristics verses frequency graph of loaded and reference antennas are illustrated in Fig.5. It is very prominent from the input impedance plot that the capacitive reactance of reference antenna below resonant frequency is nullified by the inductive loading of



(a)



(b)

Fig.6.Measured and simulated radiation characteristics of reference antenna loaded with copper coated circular dielectric material at resonance frequency 1.66 GHz (a) E-Plane, (b) H-Plane

copper coated dielectric material at resonant frequency 1.66 GHz. From the Fig.5, it is also evident that at the resonant frequencies (1.66 GHz and 2.86 GHz), the values of resistance are close to 50 ohm which lead to better matching in return loss characteristics.

The radiation characteristics of the reference antenna loaded with copper coated circular dielectric material antenna topology are defined as the E-plane (at $\varphi = 0^\circ$) and the H-plane (at $\varphi = 90^\circ$) which are shown in Fig. 6.(a) and (b). It is also observed that the simulated and measured values are agreeing very well. The cross-pol of H-plane is touching – 18 dB. The simulated and measured values of efficiency are 51.36% and 48.17% respectively. The efficiency of antenna is measured using Wheeler's cap method [13].

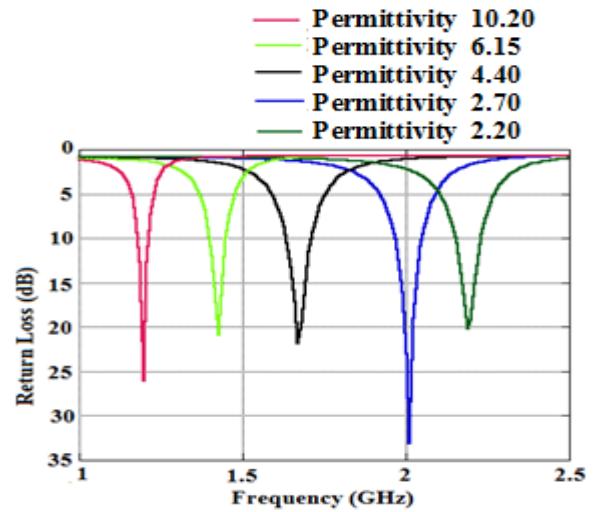


Fig.7. Parametric studies on various copper coated dielectric material keeping other parameter unchanged.

The various copper coated dielectric materials are used to examine how the permittivity affects the performance of proposed antenna topology. In this case, we also used same copper coated dielectric dimensions with a hole in centre of copper coated material. The copper coated dielectric material with permittivity 10.20 exhibits maximum miniaturization (58.39%) whereas copper coated dielectric material with permittivity 2.20 performs lowest miniaturization comparison with the reference antenna which is 23.77%. The return loss of this parametric study is shown in Fig. 7 and all performances are listed in table 1. From the table, we can draw a conclusion that miniaturization is directly proportional to the permittivity of the dielectric material.

Table1: Parametric studies on various copper coated dielectric materials keeping other parameter unchanged.

Permittivity	Resonate Frequency (GHz)	% of Miniaturization
2.20	2.18	23.77
2.70	2.01	29.72
4.40	1.66	41.95
6.15	1.42	50.34
10.20	1.19	58.39

IV. CONCLUSIONS

In this communication, we have presented a novel miniaturization technique for slot antenna. Inductive loading offered by copper coated circular dielectric material is the main reason to reduce the resonant frequency. In this technique, we get 41.95% miniaturization in resonant frequency compare with reference antenna. It has been found from a parametric study on various copper coated dielectric material that miniaturization can be improved significantly with the increment of permittivity of dielectric material. This antenna can be very useful for various wireless communication systems.

REFERENCES

- [1] SK. M. Haque, K. M. Parvez —Slot antenna miniaturization using Slit ,Strip and Loop loading Techniques,” *IEEE Trans. Antennas Propag.*, vol. 65, no. 5, pp. 2215- 2221, May 2017
- [2] M.U. Khan, M.S.Sharawi, R.Mittra,’Microstrip patch antenna miniaturisation techniques: a review,’ *IET Microw, Antennas & Propag.*, Vol. 9, Iss. 9, pp. 913–922, 2015
- [3] Y. Tawk, M.Chahoud, M.Fadous, J. Costantine, and C. G. Christodoulou,”The miniaturization of a partially 3-D printed quadrifilar helix antenna,” *IEEE Trans. Antennas Propag.*, vol. 65, no. 10, pp. 5043-5051, Oct. 2017.
- [4] SK. M. Haque, K. M. Parvez “Slot Antenna Miniaturization with Equal Electrical Path Length Using Different Shape of Loops”, *International conference on SPCOM 2018, IISc Bangalore* July 16-19, 2018.
- [5] B. Ghosh, SK. M. Haque and D. Mitra, “Miniaturization of slot antennas using slit and strip loading,” *IEEE Trans. Antennas Propag.*, vol. 59, no. 10, pp. 3922- 3927, Oct. 2011
- [6] Luukkonen, P.Ikonen, and S. Tretyakov,” Microstrip antenna miniaturization using partial dielectric material filling ”, *Microwave Opt Technol Lett*, Vol.49, Issue 1, pp. 155-159, Jan, 2007
- [7] B. Ghosh,SK M. Haque, D.Mitra and S. Ghosh “A loop loading technique for the miniaturization of non-planar and planar antennas,” *IEEE Trans AntennasPropag*, vol.58,no.6,pp 2116-2121, Jun 2010
- [8] S. Jahani, J.R. Mohassel, and M.Shahabadi, “Miniaturization of Circular Patch Antennas Using MNG Metamaterials”, *IEEE Antennas Wireless Propagation Letter*, vol. 9, pp. 1194-1196, Dec. 2010.
- [9] K.M .Parvez, SK M Haque, “Antenna miniaturization using loop and dielectric resonator”, *IEEE RFID-TA 2017*,Warsaw,Poland. Sept22-24,2017
- [10] R. Azadegan and K.Sarabandi, “A novel approach for miniaturization of slot antennas,” *IEEE Trans. Antennas Propag.*, vol. 51, pp. 421–429, Mar 2003.
- [11] N.Behdad and K. Sarabandi, “Bandwidth enhancement and further size reduction of a class of miniaturized slot antennas,” *IEEE Trans. Antennas Propag.*, vol. 52, pp. 1928–1935, Aug. 2004
- [12] HFSS ver .19.2, Ansys Corporation, Pittsburgh
- [13] D. M. Pozar and B. Kaufman, “Comparison of three methods for the measurement of printed antenna efficiency,” *IEEE Trans. AntennasPropag.*, vol. 36, pp. 136–139, Jan. 1988

Quantum Random Number Generator with One and Two Entropy Sources

Gautam Shaw

Department of Electrical Engineering

IIT Madras

Chennai, India

ee15d047@ee.iitm.ac.in

Sivaram SR

Department of Electrical Engineering

IIT Madras

Chennai, India

sivaram01@live.com

Anil Prabhakar

Department of Electrical Engineering

IIT Madras

Chennai, India

anilpr@iitm.ac.in

Abstract—Quantum random number generators (QRNGs) are an integral part of quantum key distribution (QKD) systems. To better understand the inherent physical processes, we compare the random numbers generated by two separate schemes, one is based on entropy (arrival time of photons) and another with an additional source of entropy (space) i.e., path superposition of arrival time of photons from a weak coherent source on a gated InGaAs single photon detector.

Both experiments yield bits that appear random. However, they satisfy different criteria of randomness. The weak coherent source has a Poissonian distribution and extracting the variation about the arrival time of photons on gated SPD yields a source of random numbers that pass most of the Dieharder Tests. With the inclusion of superposition, we obtain random numbers that pass all the Dieharder tests. The physical origins of the random numbers in the two experiments is different, one is single entropy source based and other one is two entropy source based, and this is reflected in the outcomes of the different tests for randomness.

I. INTRODUCTION

Random numbers have played essential roles in a wide variety of fields, such as cryptography, scientific simulations, random walks. There are two methods to generate a random number sequence namely pseudo-random number generation (PRNG) and true random number generation (TRNG). A pseudo random number generator is a mathematical formula, or more generally a deterministic algorithm which, starting from a certain initial number (seed) that defines the initial state, produces a string of numbers that looks random in the sense that it possesses a certain set of desirable statistical properties, but in fact it is completely predictable and highly losslessly compressible by definition [1]. Recently, physical random number generation techniques were reported in literature [2] [3], which are based on chaotic behaviour of a semiconductor laser. Generally speaking, the above schemes are not able to generate true random numbers with information-theoretically provable randomness. A quantum random number generator, generate bits from the fundamentally probabilistic nature of quantum processes [4] [5]. Examples of demonstrated QRNG include two path splitting of single photon [6], time of generation or counting of photons [7] [8], fluctuations of the vacuum state using homodyne detection techniques

This work was supported by Ministry of Human Resources and Development (MHRD) under order number 35-8/2017-TS.

[9], photon number path entangled state [10], as well as interferometric schemes [11]. Another unique QRNG was demonstrated which is based on random population of the spatial modes of a beam splitter when both inputs are simultaneously fed with the indistinguishable weak coherent pulses [12]. Furthermore, most of the QRNGs uniformly distributed random numbers, are used for quantum cryptography. Apart from that many QRNGs having certain non-uniform distributions, or other tailored statistical properties, are used for Monte Carlo simulations and coherent state quantum key distribution. However, the use of a quantum random number generator (QRNG) seems to be crucial in quantum key generation, used for encrypted data communication in optical fiber or free space [13] [14]. In this paper, we report on two methods of quantum random number generation, one is based on a single source of entropy i.e., arrival time of photon from weak coherent source and other one is based on two sources of entropy i.e., superposition of paths with uncertainty in arrival time of photon to single photon detector (SPD). We have converted path superposition into time superposition and hence we used only one SPD instead of two. Both methods have passed standard ENT, NIST test and the more exhaustive Dieharder tests. Comparative analysis of all standard statistical tests are reported for both schemes.

II. RANDOMNESS IN ARRIVAL TIME

In quantum optics, attenuated pulses with mean photon number μ , follow a Poissonian distribution.

$$P(n) = \frac{e^{-\mu} \mu^n}{n!} \quad (1)$$

where $P(n)$ is the Poisson distribution with mean photon number μ . A weak coherent pulse (WCP), with $\mu \sim 0.1$, is loosely interpreted as having one photon present in 1 out of 10 pulses, but within the coherence time of the source. However, the exact time of emission remain indeterminate. The photon is incident on a single photon detector (SPD), with a weak laser source, hence the probability of detection of a single photon depends on the measurement position (gated by a time bin) within the coherence time. Each detection event in the Poissonian distribution is independent of the past distribution, hence uncertainty in arrival time of photon added

with uncertainty in detection (due to detector efficiency) cause random events of photon detection. These random events are collected as time stamp using a time stamp module. If the photon detection event occurs in an even multiple of the time stamp clock, the bit generated is '0' else the bit generated is '1'.

A. Experimental set-up

Optical pulses having a width of 2 ns, generated from a small form pluggable (SFP) module, at a period of 64 ns are attenuated to achieve $\mu \sim 0.1$. RF pulses in synchronization with attenuated optical pulses are used for external gating of a SPD. Photon arrival times are recorded by using a TIVA microcontroller board. SPD settings are kept at the lowest noise mode, $T_{\text{Hold}} = 10 \mu\text{s}$, $V_{\text{ex}} = 2.5 \text{ V}$, Temperature = -40°C . The internal clock of the same TIVA board was set to 80 MHz for time stamp collection. The RF delay controller was used to synchronize the external gate with the arrival time of the WCP at the SPD. The gate delay was swept from 0 to 32 ns in steps of 0.5 ns and the photon counts per second were recorded. Fig. 1 shows the set-up used for this experiment, and Fig. 2 shows the peak in counts/s obtained when the gate delay was correctly synchronized to the arrival of the WCP at the SPD, at 3.5 ns.

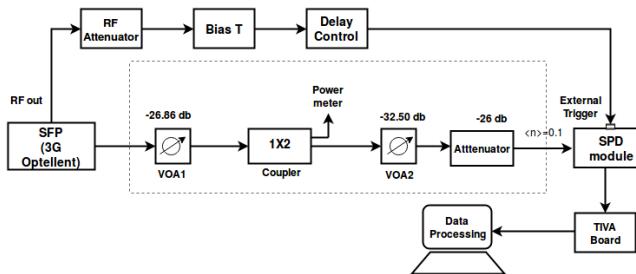


Fig. 1. Experimental set-up for QRNG based on arrival time of photon.

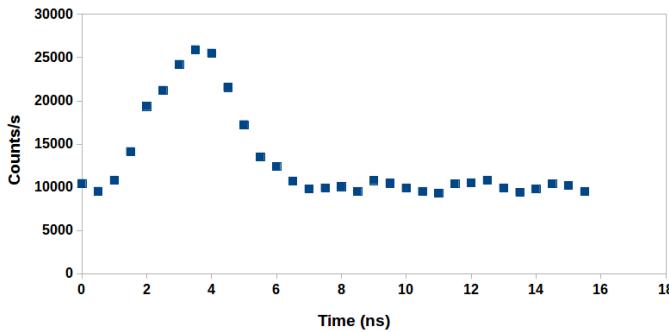


Fig. 2. Maximum counts/s were observed at RF delay of 3.5 ns.

III. RANDOMNESS FROM UNCERTAINTY IN TIME AND SPACE

To improve upon the entropic nature of the random number generator, we considered the addition of a beam splitter before

the SPD in Fig. 1. This modification is shown schematically in Fig. 3. A single photon would chose one of two different paths, and this would add an additional randomness to the photon arrival times described in Sec. II. We convert this spatial uncertainty to temporal uncertainty by adding a delay line on one path, and then combining the two paths with a 2x1 directional coupler, before feeding it to the SPD. The disadvantage of this setup is that we effectively lose half the photons in the coupler, since a 2x1 coupler actually has a dark port that is neglected. More complex optical constructs, using polarization beam splitters, would eliminate this loss, but is not necessary while seeking to build random number generators.

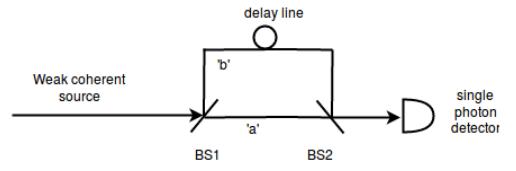


Fig. 3. Conversion of spatial superposition into temporal superposition

A. Experimental set-up

We modified the experimental set-up described in Fig. 1 by adding two 3 dB 1x2 couplers with a 20 ns delay line at the end of quantum channel, shown schematically in Fig. 4. To accommodate this modification, the pattern of external gate pulses were also suitably modified. In a period of 64 ns, 2 RF pulses were generated using the Optellent unit and an RF coaxial delay line was used to match with the timing of photon arrival to the SPD through paths 'a' and 'b'. Synchronization was done using measurement of counts/s by varying the gate delay using a delay controller as shown in Fig. 4. The photon arrival times were again recorded by the time stamp module, written on the TIVA-C LaunchPad. The SPD settings were retained at the lowest noise mode, $T_{\text{Hold}} = 10 \mu\text{s}$, $V_{\text{ex}} = 2.5 \text{ V}$, Temperature = -40°C . Experiment for each scheme was run for 72 hours to collect 1 Gb of binary data. A simple modulo (time stamp, 2) gives the binary values '0' or '1' based on the arrival time of photon in either even or odd time slots, respectively. The collected data was processed and converted to binary files, passed through ENT, NIST and Dieharder tests.

IV. RANDOMNESS TEST

We have chosen three batteries of statistical tests to evaluate the two different schemes of QRNG, namely the ENT, NIST and DIEHARD suite of tests, considered by most to be sufficient in qualifying the device for its use in cryptography. ENT is a series of basic statistical tests which evaluate the random sequence in some of the elementary features such as equal probability of ones and zeros and serial correlation.

- Entropy test - Checks the entropy per bit. For a completely random bit stream this approaches 1.

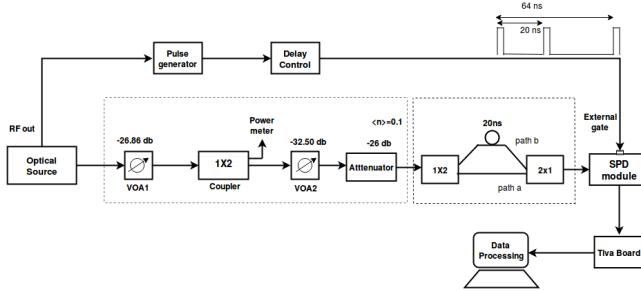


Fig. 4. Experimental set-up for QRNG based on spatio-temporal distributions of photons.

TABLE I
ENT TEST RESULTS

Test	Result 1	Result 2
Entropy per bit	1.0000	1.0000
Compression ratio	0.0	0.0
Chi Square value	0.34	0.86
Mean Value Bits	0.5000	0.5000
Monte Carlo value of "pi" error	0.11	0.01
Serial Correlation Coefficient	0.003508	0.000452

- Compressibility test - Check by how much the bit stream file can be compressed. For a completely random bit stream this will be zero.
- Chi Square test - It is also called a "goodness of fit" statistic test, because it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.
- Mean value of a Bit - Check the proportion of Zeros and Ones in the random number stream. For completely random bit stream this should be near 0.5
- Monte Carlo Test - This test calculates the value of "pi" using Monte Carlo method using the random bit stream and gives the error in the calculated value. The error value should be as close to zero as possible.
- Serial Correlation Coefficient - Should be as close to zero as possible for completely random data.

To further exploit some subtle imperfections hidden in both QRNG schemes, we test the sample sequence of 100 Mb from both schemes using NIST suite. Results-1 and Results-2 are ENT test values for two different QRNG schemes based on arrival time and spatial distributed arrival time respectively. Monte carlo value of error is 0.11% in single entropy based scheme, and for bi-entropy based scheme, it is 0.01%. Serial Correlation Coefficient is of the order of 10^{-3} and 10^{-4} in Scheme 1 and Scheme 2 respectively. Although both are uncorrelated, Scheme 2 is an order better than Scheme 1.

Table II shows the NIST test p-values for both QRNG schemes. After analysis of statistical test values by both the schemes, the minimum pass rate for each statistical test is approximately 96 for a sample size of 100 binary sequences in our QRNG Scheme-1, but the pass rate is 22 in Runs

TABLE II
NIST TEST RESULTS

Test	P-value, success rate	P-value, success rate
Frequency	0.935716, 99/100	0.474986, 100/100
Block Frequency	0.000089, 97/100	0.191687, 99/100
Cumulative Sums	0.191687, 99/100	0.595549, 100/100
Runs	0.000000, 22/100	0.000026, 95/100
Longest Run	0.739918, 99/100	0.719747, 100/100
Rank	0.213309, 99/100	0.971699, 99/100
FFT	0.883171, 100/100	0.275709, 96/100
Approximate Entropy	0.534146, 99/100	0.108791, 99/100
LinearComplexity	0.883171, 98/100	0.494392, 97/100
Non overlapping	0.213309, 100/100	0.236810, 100/100
Serial test	0.437274, 99/100	0.315837, 100/100
Universal	0.883171, 98/100	0.058984, 100/100

test for Scheme 1. Similarly, the pass rate is about 96 for a sample size of 100 binary sequences in photon arrival with path superposition based QRNG Scheme-2, also pass rate is 95 in Runs test for a sample size of 100 binary sequences. The p-value corresponding to Frequency test is 0.935716 and 0.474986 for Scheme-1 and Scheme-2 respectively, which are the measure of our confidence that we have equal numbers of '1's and '0' across the 1000 bit sequence. The p-value is calculated from the χ^2 data. Each NIST parameter is observed for the entire bit sequence together and its χ^2 is computed against the theoretical values obtained for the same parameter considering a corresponding bit sequence under the assumption of randomness, and finally the observations are converted into p-value under the assumption of standard normal distribution.

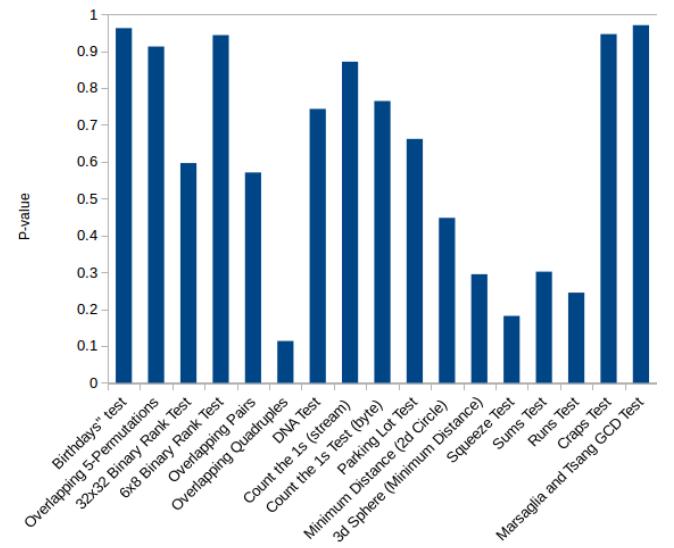


Fig. 5. Dieharder test results for QRNG with scheme-1(single entropy: arrival time of photon on SPD)

If the variation in χ^2 is large, then the p-value becomes smaller, indicating that the bit sequence under consideration tends towards being non-random. Scheme 1 failed the Runs

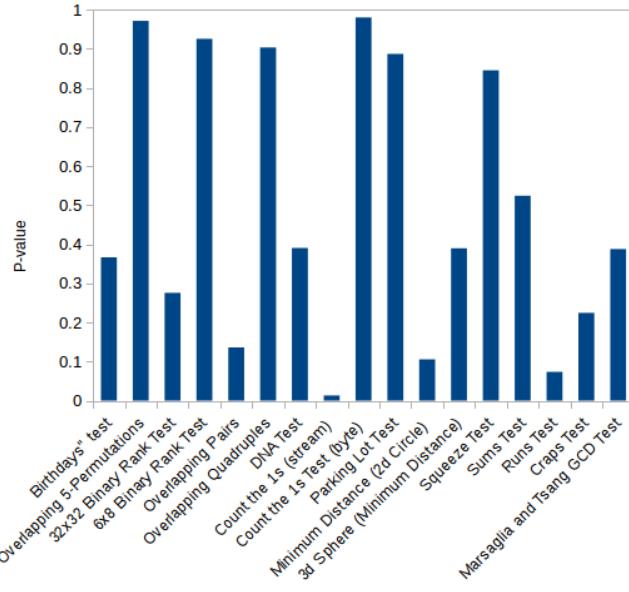


Fig. 6. Dieharder test results for QRNG with Scheme-2 (bi-entropy: arrival time of photon with path superposition)

test because of a high deviation of its distribution from the normal distribution for 100 bit streams. Except for the Cumulative sum test and Rank test, we can see that p-values of other individual test for Scheme 1 are either equal or greater than than p-values of Scheme 2. It shows that Chi Square distribution is closer to a Gaussian distribution and that the tests for scheme 2 are more consistent with that for a true random sequence. The Dieharder Test Suite, is by far, regarded as the best available standard of random number testing suite in the industry. In addition to George Marsaglias Diehard Tests, which are a series of statistical tests that were developed by him and published in 1995, the Dieharder test suite, comprises of numerous statistical tests from other sources too, including some of the tests of the NIST Test Suite. Dieharder results for both schemes are shown in Fig. 5 and Fig. 6. In most cases, the obtained p-values are above 0.01 and below 0.99 and only few exceptions are obtained outside these boundaries, producing a warning messages. However, the number of suspect p-values is within the expected statistical failure rate, assuming a 1% significance level, thus perfectly normal for a true random number generator, as explained in the Dieharder documentation.

V. CONCLUSION

We have demonstrated two different schemes for quantum random number generation, using a weak coherent pulse as the source of photons. While the first scheme is based on only the arrival time of photons, the second scheme is based on arrival time and path superposition of photons. Both schemes have successfully passed all statistical tests, but with different levels of confidence (p-values). We observe that having more than

one source of entropy offers a higher degree of randomness in the output bit stream, although both schemes are likely to be suitable candidates for use in quantum key distribution.

VI. ACKNOWLEDGMENT

This work was supported by Ministry of Human Resources and Development (MHRD) under order number 35-8/2017-TS. We are thankful to Dr. Prabha Mandayam, Shashank Ranu and Shaik Basharath at IIT Madras, and Dr. Anindita Banerjee, QuNu labs, for many conversations on QKD and QRNG.

REFERENCES

- [1] P. Hellekalek, "Good random number generators are (not so) easy to find," *Mathematics and Computers in Simulation*, vol. 46, no. 5-6, pp. 485–505, 1998.
- [2] A. Uchida, K. Amano, M. Inoue, K. Hirano, S. Naito, H. Someya, I. Oowada, T. Kurashige, M. Shiki, S. Yoshimori *et al.*, "Fast physical random bit generation with chaotic semiconductor lasers," *Nature Photonics*, vol. 2, no. 12, p. 728, 2008.
- [3] I. Reidler, Y. Aviad, M. Rosenbluh, and I. Kanter, "Ultrahigh-speed random number generation based on a chaotic semiconductor laser," *Physical review letters*, vol. 103, no. 2, p. 024102, 2009.
- [4] F. Xu, B. Qi, X. Ma, H. Xu, H. Zheng, and H.-K. Lo, "Ultrafast quantum random number generation based on quantum phase fluctuations," *Optics express*, vol. 20, no. 11, pp. 12 366–12 377, 2012.
- [5] W. Wei and H. Guo, "Bias-free true random-number generator," *Optics letters*, vol. 34, no. 12, pp. 1876–1878, 2009.
- [6] T. Jennewein, U. Achleitner, G. Weihs, H. Weinfurter, and A. Zeilinger, "A fast and compact quantum random number generator," *Review of Scientific Instruments*, vol. 71, no. 4, pp. 1675–1680, 2000.
- [7] M. Stipčević and B. M. Rogina, "Quantum random number generator based on photonic emission in semiconductors," *Review of scientific instruments*, vol. 78, no. 4, p. 045104, 2007.
- [8] P. Bronner, A. Strunz, C. Silberhorn, and J.-P. Meyn, "Demonstrating quantum random with single photons," *European journal of physics*, vol. 30, no. 5, p. 1189, 2009.
- [9] C. Gabriel, C. Wittmann, D. Sych, R. Dong, W. Mauerer, U. L. Andersen, C. Marquardt, and G. Leuchs, "A generator for unique quantum random numbers based on vacuum states," *Nature Photonics*, vol. 4, no. 10, p. 711, 2010.
- [10] O. Kwon, Y.-W. Cho, and Y.-H. Kim, "Quantum random number generator using photon-number path entanglement," *Applied Optics*, vol. 48, no. 9, pp. 1774–1778, 2009.
- [11] B. Qi, Y.-M. Chi, H.-K. Lo, and L. Qian, "High-speed quantum random number generation by measuring phase noise of a single-mode laser," *Optics letters*, vol. 35, no. 3, pp. 312–314, 2010.
- [12] T. F. da Silva, G. Xavier, G. Amaral, G. Temporão, and J. von der Weid, "Quantum random number generation enhanced by weak-coherent states interference," *Optics express*, vol. 24, no. 17, pp. 19 574–19 580, 2016.
- [13] E. Diamanti, H.-K. Lo, B. Qi, and Z. Yuan, "Practical challenges in quantum key distribution," *npj Quantum Information*, vol. 2, p. 16025, 2016.
- [14] S. K. Ranu, G. K. Shaw, A. Prabhakar, and P. Mandayam, "Security with 3-pulse differential phase shift quantum key distribution," in *2017 IEEE Workshop on Recent Advances in Photonics (WRAP)*. IEEE, 2017, pp. 1–7.

Simulation of Emission Wavelength of Quantum Dot Based Single Photon Sources

Jyothish M

Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India
ee15d010@ee.iitm.ac.in
and

IISU, Vikram Sarabhai Space Centre
Indian Space Research Organisation
Thiruvananthapuram, India

Fredy Francis

Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India
ee14d020@ee.iitm.ac.in

R. Manivasakan

Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai, India
rmani@ee.iitm.ac.in

Abstract—Advances in quantum information processing and the requirements of quantum key distribution schemes have made high quality single photon sources, extremely essential. Here generation of a single photon from a semiconductor quantum dot using a semi-classical approach is investigated. Finite Element Method is used for the Eigen mode analysis of a typical pyramidal semiconductor quantum dot [3]. A design methodology is also proposed for obtaining the required emission wavelengths. Additionally, effects of wetting layer, height to base ratio and strain due to lattice mismatch are investigated. An Empirical relationship is obtained between pyramid geometry and emission wavelength. The simulation results were verified against the experimental works including [7] in 1.2 μm to 1.3 μm emission regime, and good match was observed.

Index Terms—Single Photon Sources, Quantum Dot, FEM simulation, Emission wavelength selection

I. INTRODUCTION

Quantum information science have seen massive growth in the last few decades. From qubits to random number generators to quantum key distribution, Single Photon Sources (SPS) play an extremely important role. These have paved its way for an active research in fabrication and characterization of these single photon sources. A number of techniques have been proposed and experimentally demonstrated by different research groups around the globe.

Single photon generators can be broadly and loosely divided into deterministic and probabilistic. The former being 'on-demand', where single a photon can be emitted whenever the user wishes whereas in the latter generation is probabilistic and the presence of the heralded photon (created in pair) is heralded by the presence of heralding photon. Deterministic sources are usually based on single atoms, ions or molecules; atomic ensembles, quantum dots (QD), color centers etc and probabilistic sources are obtained using Parametric Down Conversion (PDC), Four Wave Mixing (FWM) etc [4].

SPS using QD is attractive as it can be fabricated cheaply using well established growing techniques such as Stranski-Krastanov, where molecular beam epitaxy is used to create

tiny islands of small Band Gap (BG) semiconductor on large BG ones using self-assembly [4]. Distributed Bragg Reflector (DBR) can be used to control the direction of photon emissions, also they can be integrated into cavities like disk, micro-pillar, sphere, photonic crystal [4]. This helps in the increasing the emission efficiency of the QD by *Purcell effect*, if the energy and polarization of the emission field matched to that of the cavity. Also the cavities help in collecting the single photons in a single spatial mode. A downside of QD based SPS is that the second order coherence function ($g^{(2)}(0)$) is still not as low as other competing technologies for QD.

In this work, we simulate a InAs/GaAs pyramidal Quantum Dot (QD), study its optical properties and try to figure out the single photon emission wavelength based on the QD physical parameters. Finite Element Method (FEM) is used for this and the effect of QD size on the emission wavelength is studied. There have been a number of studies in the spectral and dynamic properties of QD Lasers. QD material, form, structure and properties have been very widely investigated owing to their numerous advantages as a laser like, lower threshold current density, high differential gain, stability of threshold current to temperature variations, wide modulation bandwidth, direct modulation without frequency chirping [2].

The S-K mode growth dynamics for the QD and its corresponding sizes and shapes were investigated in [6]. [8] investigates the effect of InAs QD position relative to the InGaAs Strain Reducing Layer (SRL). The SRL (InGaAs) helps to relieve the stain hence modifying confinement potential, which allows manipulation of transition Energies. Optical transition energy of the QD was obtained by solving 3D Schrodinger equation using FEM while considering the strain due to lattice mismatch. The InAs QD shape was assumed to be that of a lens, with GaAs as the Matrix and using InGaAs as the SRL. It was found that InGaAS in the cap layer increases the QD size whereas when its below the QD size is significantly smaller. [3] uses low temperature, high rate GaAs coating of InAs QD to obtain non-truncated pyramidal InGaAS QDs in GaAs. Then Transmission Electron Microscopy (TEM)

was used to study the mechanics and composition of the structure. They could develop models of elastic strain and stress, chemical distribution, 3D shape and stresses of matrix, wetting layer and the QD. Finite Element Method (FEM) was used to model the pyramidal quantum dots and to study about emission wavelength of such structures. Effect of strain due to lattice mismatch on energy bandgap is also accounted for. [5] extends the numerical model of S-K QD growth and uses Photoluminescence (PL) measurements to investigate the buried dot size. Schrodinger equation was solved for one band, single particle scenario using effective mass approximation to estimate the size and aspect ration of buried QD for different growth rates.

II. THEORETICAL BACKGROUND

A. Structure of Semiconductor Quantum Dots

By confining charge carriers of a semiconductor within dimensions comparable with their de Broglie wavelength, atomic-like sharp peaked energy levels and comparable emission spectrum can be obtained. InAs grown over GaAs substrate follows Stranski Krastanov (SK) growth mechanism which is one among the three primary modes of epitaxial growth of thin films over crystal surface. In SK growth, adsorbate form several monolayers over the substrate before islands are created owing to strain build up and chemical potential of the deposited film [13]. Island formation can be either dislocated or coherent. Dislocation free, coherent islands can be formed by introducing undulations in the layers nearer to the substrate which relieves strain and helps in matching the lattice constants of wetting layer and island to that of bulk [14]. This ability to alter the onset of SK growth enables control over the geometry and size which in turn can alter the optical properties. Island geometry can also be controlled by controlling the growth rate and substrate relief [15] [16].

B. Optical properties of single quantum dots

A particularly successful growth technique of QD is the self organized growth of InAs over GaAs substrate. The InAs/GaAs QD shows very good optical properties. Overgrowth of InAs on GaAs beyond a critical thickness (wetting layer thickness) causes island of InAs with pyramidal shape to form over the GaAs substrate. The problem of random spatial positions of QDs may be overcome by nano sized pits on growth surface which favour QDs to grow. As InAs has lower electronic bandgap compared to GaAs substrate, it trap electrons and holes. A biexciton state can be created by illuminating the QD with a picosecond laser. One of the trapped electron recombines with a hole and creates a photon (biexciton photon, X_2), the e-h pair left over also combines later on to create the exciton Photon, (X).

C. Mathematical Model

1) Schroedinger wave equation (SWE) : The time-dependent Schroedinger wave equation is given by [9],

$$i\hbar \frac{\partial}{\partial t} \Psi(r, t) = \hat{H} \Psi(r, t) \quad (1)$$

Where $\Psi(r, t)$ is the wave-function, r is the spatial position at time t . \hat{H} is the Hamiltonian operator of the system. Non-relativistic Schrodinger equation for a partial moving in a potential field is a famous example.

$$i\hbar \frac{\partial}{\partial t} \Psi(r, t) = \left[-\frac{\hbar^2}{2m} \nabla^2 + V(r, t) \right] \Psi(r, t) \quad (2)$$

Wave function can form so called stationary states which are standing wave patterns. The stationary states can be described by time-independent Schrodinger wave equation.

$$\hat{H} \Psi(r) = \hat{E} \Psi(r) \quad (3)$$

Where \hat{E} is the energy operator. Time-indipendent (TI) Schrodinger wave equation for a single particle evolving in a potential field is an Eigen value problem with eigen values corresponding to energy levels of the quantum system.

$$\left[-\frac{\hbar^2}{2m} \nabla^2 + V(r) \right] \Psi(r) = E \Psi(r) \quad (4)$$

Using effective mass approximation [10], an electron/hole moving in a semiconductor can be described with Schrodinger wave equation with a fictitious mass called effective mass m_{eff} .

D. Computational Methods: FEM

An FEM simulation was performed to calculate and visualize the eigen wave-functions. The simulation domain is divide into a large number of tetrahedron elements. The system matrix is obtained by joining individual element matrices and the eigen decomposition is performed. The result is exported to '.vtk' format and visualized using a software (Paraview).

A typical FEM simulation has five steps. First step is to divide the complex computational domain into a collection of simple sub domains. Secondly, the partial differential equations being solved are represented by a set of linear equations within each individual elements. Third step is to combine element equations into a global system of equation. Fourth step is to solve the global system of equations thus formed using standard methods and fifth step is to calculate field at each point within element by element wise interpolation.

The time-independent Schrodinger wave equation TI-SWE dot is given by 4 as,

$$\left[-\frac{\hbar^2}{2m_{\text{eff}}} \nabla^2 + V(r) \right] \Psi(r) = E \Psi(r) \quad (5)$$

Here $m_{\text{eff}} = m_0 m_r$. Where m_0 is the mass of a free electron and m_r is taken from table I.

For computational convenience we convert the 5 into the form,

$$\left[\frac{\gamma}{m_r} \nabla^2 + V(r) \right] \Psi(r) = E \Psi(r) \quad (6)$$

Where Laplacian ∇^2 is specified in nm^{-2} , $V(r)$ and $E(r)$ are in eV.

$$\gamma = -\frac{\hbar^2}{2m_0} \frac{1}{\beta p^2}$$

Here, β is the conversion factor from eV to Jules ($\approx 1.6022 \times 10^{-19}$) and p is the space scaling factor to convert nanometers to meters ($p = 10^{-9}$).

1) *FEM Simulation Procedure*: Galerkin FEM formalism [11] is used for simulation here, which involves equating the integral of inner products between weight functions and residuals to zero. The weight functions are the basis functions in Galerkin approach. If we consider a tetrahedral finite element with four node and with basis functions $\{M_j\}$, the field at any point within the element can be represented as a linear interpolation of field at node values as,

$$\Psi(r) = \sum_i M_i \Psi_i \quad (7)$$

Deriving weak form of TI Schrdinger (6),

$$\sum_j \int_{\Omega} M_j \left(\frac{\gamma}{m_r} \nabla^2 + V(r) - E \right) \Psi(r) d\Omega = 0 \quad (8)$$

Substituting (7) and assuming potential V is constant within an element,

$$\begin{aligned} & \frac{\gamma}{m_r} \sum_j \sum_i \int_{\Omega} M_j \nabla^2 M_i d\Omega \Psi_i \\ & + V \sum_j \sum_i \int_{\Omega} M_j M_i d\Omega \Psi_i = E \sum_j \sum_i \int_{\Omega} M_j M_i d\Omega \Psi_i \end{aligned} \quad (9)$$

From Green's theorem,

$$\int_{\Omega} \nabla M_j \cdot \nabla M_i d\Omega + \int_{\Omega} M_j \nabla^2 M_i d\Omega = \int_{\Gamma} M_j \frac{\partial M_i}{\partial n} d\Gamma \quad (10)$$

Where the last term $\frac{\partial M_i}{\partial n}$ represents the change of basis function M_i along the outward normal n to the enclosing surface Γ to the element volume Ω . The last term can be equated to zero as we assume zero field for periphery of computational domain and for the internal elements this term cancels out as we combine the element metrics to obtain the global system matrix. So we have,

$$\int_{\Omega} M_j \nabla^2 M_i d\Omega = - \int_{\Omega} \nabla M_j \cdot \nabla M_i d\Omega$$

Substituting in (9),

$$\begin{aligned} & -\frac{\gamma}{m_r} \sum_j \sum_i \int_{\Omega} \nabla M_j \cdot \nabla M_i d\Omega \Psi_i + V \sum_j \sum_i \int_{\Omega} M_j M_i d\Omega \Psi_i \\ & = E \sum_j \sum_i \int_{\Omega} M_j M_i d\Omega \Psi_i \end{aligned} \quad (11)$$

In order to simplify the integration process, we use a reference tetrahedron in natural coordinate system as shown in Fig. 1b. A point in reference space (ξ, η, ζ) can be converted to a point in real space using the equation,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + J \begin{bmatrix} \xi \\ \eta \\ \zeta \end{bmatrix} \quad (12)$$

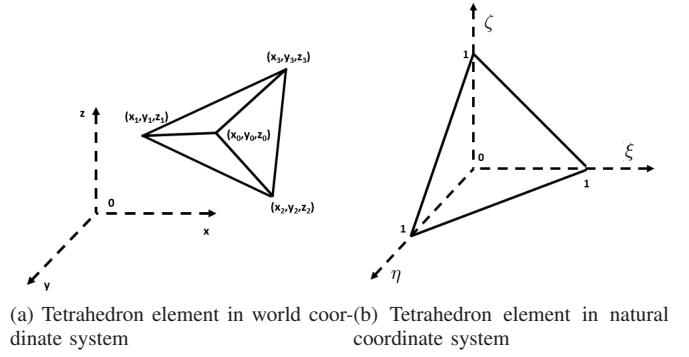


Fig. 1: Tetrahedron element

Where (x_0, y_0, z_0) , (x_1, y_1, z_1) , (x_2, y_2, z_2) and (x_3, y_3, z_3) are the corners of the real tetrahedron and the Jacobian J is given by

$$J = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \zeta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} & \frac{\partial y}{\partial \zeta} \\ \frac{\partial z}{\partial \xi} & \frac{\partial z}{\partial \eta} & \frac{\partial z}{\partial \zeta} \end{bmatrix} = \begin{bmatrix} x_1 - x_0 & x_2 - x_0 & x_3 - x_0 \\ y_1 - y_0 & y_2 - y_0 & y_3 - y_0 \\ z_1 - z_0 & z_2 - z_0 & z_3 - z_0 \end{bmatrix} \quad (13)$$

The basis vectors used in natural coordinate system are $\{N_i\}$.

$$\begin{bmatrix} N_0 \\ N_1 \\ N_2 \\ N_3 \end{bmatrix} = \begin{bmatrix} 1 - \xi - \eta - \zeta \\ \xi \\ \eta \\ \zeta \end{bmatrix}$$

Then,

$$\int_{\Omega} \nabla M_j \cdot \nabla M_i d\Omega = \int_{\Omega_n} \nabla N_j \cdot \nabla N_i \det(J) d\Omega_n \quad (14)$$

and

$$\int_{\Omega} M_j M_i d\Omega = \int_{\Omega_n} N_j N_i \det(J) d\Omega_n \quad (15)$$

Where Ω_n is the reference element volume, $\det(J)$ is the determinant of Jacobian defined in (13). As the Jacobian matrix is constant for a given element, so is its determinant. It is straightforward to prove that,

$$\int_{\Omega_n} N_j N_i \det(J) d\Omega_n = \begin{cases} \det(J) \frac{2}{120} & \text{if } i=j \\ \det(J) \frac{1}{120} & \text{if } i \neq j \end{cases} \quad (16)$$

and

$$\int_{\Omega_n} \nabla N_j \cdot \nabla N_i \det(J) d\Omega_n = \det(J) \frac{1}{6} \langle J^{-1} R_i, J^{-1} R_j \rangle \quad (17)$$

where \langle , \rangle represent inner product and R_i and R_j are i^{th} and j^{th} rows of the matrix,

$$R = \begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Substituting (16) and (17) in equations (14) and (15) then substituting in (11) we can obtain the required element matrix.

$$Ax = EBx \quad (18)$$

Where x is a column vector of field (wave-function) values of four nodes of the tetrahedron. A and B are matrices given by,

$$A_{ij} = -\frac{\gamma}{m_r} \int_{\Omega_n} \nabla N_j \cdot \nabla N_i \det(J) d\Omega_n + V \int_{\Omega_n} N_j N_i \det(J) d\Omega_n$$

$$B_{ij} = \int_{\Omega_n} N_j N_i \det(J) d\Omega_n$$

2) *Assembling to global system matrix:* The global system matrix is obtained by adding the contributions from different elements carefully and mapping the local indexes into a global node index. This global system matrix is a generalized Eigen value problem, with Eigen value E and Eigen vector X , represented by equation,

$$A_g X = E B_g X \quad (19)$$

As the wave-forms are expected to vanish at the boundaries, Dirichlet boundary condition is imposed which involves removing the rows and columns of node at the boundary. This Eigen value problem is solved at the nodal points using standard methods to obtain the wave-functions.

III. APPLICATION OF FEM TO PROPOSED QUANTUM DOTS

A. Simulation Results

The simulation procedure is as follows. The physical geometry is modeled using a 3d modeler (blender software) and exported to a geometry file using a custom script. A meshing software (gmsh) imports .geo file and an unstructured mesh is created which is then exported to a .msh format. The FEM solver code developed in python reads the mesh file and performs eigen mode simulations. The eigen fields are exported to .vtu format and visualized using the software, paraview. The flow chart is given in Fig. 2. The pyramidal quantum dot model used and the mesh generated are shown in Fig. 3a and Fig. 3b respectively.

B. Eigen modes of TI Schrödinger wave equation simulated

Eigen modes obtained from the simulation are visualized using a free software called paraview. The quantum dot size is varied from 5 nm to 10 nm using a python script and the results are shown in Fig. 4.

C. Prediction of emission wavelength from quantum dot

Simulation of eigen modes of TI-SWE is performed for various sizes, height to base ratios and wetting layer thickness. Light holes cannot take part in radiate emission as their spin ($\pm 1/2$) difference between electrons ($\pm 1/2$) does not match up with the spin of photon ± 1 . The energy of a photon emitted in an electron hole recombination is,

$$E_{ph} = E_e - E_h - J_{sh} \quad (20)$$

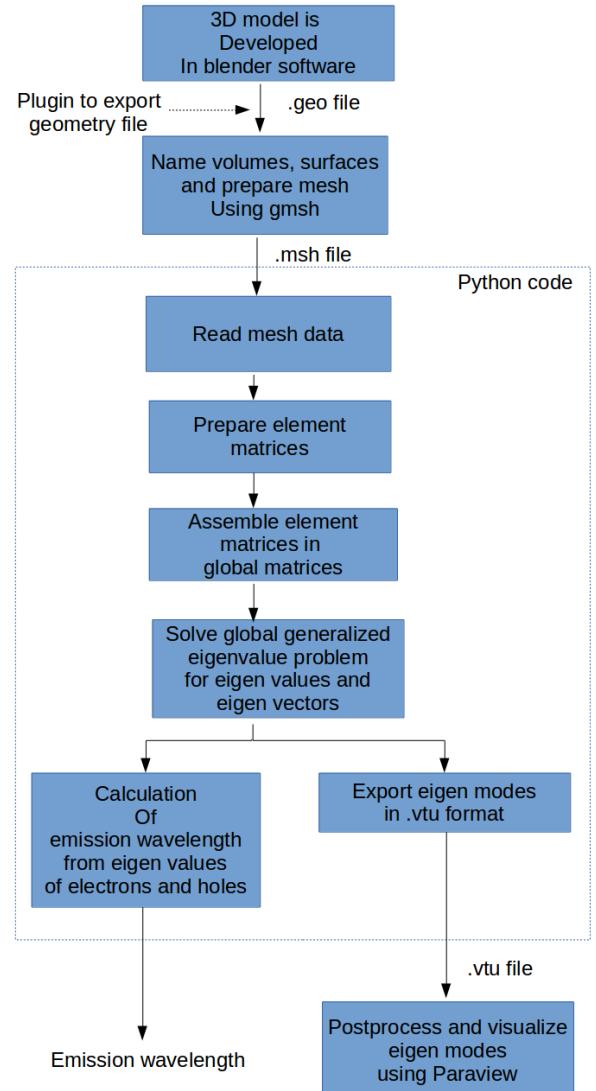
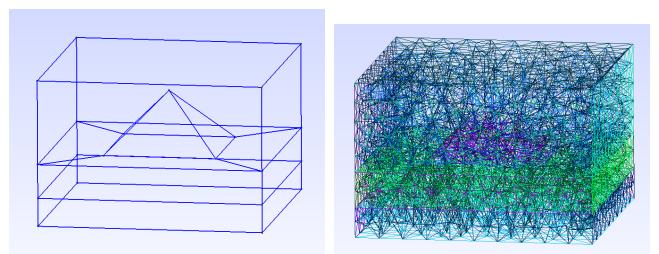


Fig. 2: Simulation procedure



(a) Gmsh model of pyramidal quantum (b) 3D mesh generated using Gmsh dot

Fig. 3: Pyramidal quantum dot model for FEM simulations

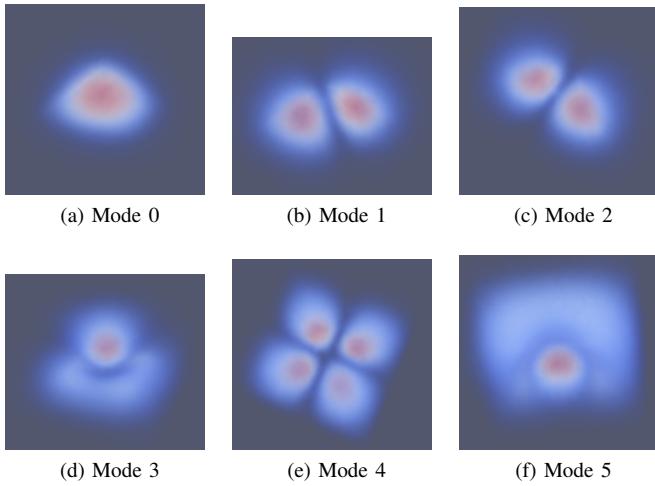


Fig. 4: Eigen solutions of Schrodinger wave equation solved using FEM technique and visualized using paraview

Where E_e is energy of an electron in valence shell, E_h is the energy of heavy hole and J_{sh} is the colombian interaction energy ($\approx 10\mu\text{eV}$). The emission wavelength may be calculated from the photon energy using the relation,

$$\lambda = \frac{hc}{E_{ph}} \quad (21)$$

Emission wavelength for various quantum dot dimensions are plotted in Fig. 5.

1) *Effect of strain on band-gap:* Strain developed due to lattice mismatch between quantum dot material InAs and substrate material GaAs changes energy bandgap considerably. The expressions given below gives the hydro-static and uniaxial strains respectively [12].

$$\begin{aligned} \epsilon_h &= \epsilon_{xx} + \epsilon_{yy} + \epsilon_{zz} \\ \epsilon_b &= \epsilon_{zz} - \frac{1}{2}(\epsilon_{xx} + \epsilon_{yy}) \end{aligned}$$

The lattice mismatch is

$$\epsilon_{xx} = \epsilon_{yy} = \frac{a_{substrate} - a_{qdot}}{a_{qdot}}$$

Here $a_{substrate}$ and a_{qdot} are lattice constants of substrate (GaAs) and quantum dot (InAs) materials respectively.

$$\epsilon_{zz} = -2 \frac{C_{12}}{C_{11}} \epsilon_{xx}$$

Change in conduction band due to strain is,

$$\delta E_c = a_c \epsilon_h$$

and change of valance band is given by,

$$\delta E_v = a_v \epsilon_h - \frac{1}{2} b \epsilon_b$$

Finally, the total change in bandgap energy is given by $\delta Eg = \delta E_c - \delta E_v$, and the net bandgap of InAs quantum dot

TABLE I: Parameters used for calculation

Material	$m_e^*(m_0)$	$m_h^*(m_0)$	$C_{11} 10^{11} (\text{dyn cm}^{-2})$	$C_{12} 10^{11} (\text{dyn cm}^{-2})$	$(a_c - a_v) (\text{eV})$	b (eV)
InAs	0.023	0.41	8.32	4.52	-4.08	-1.8
GaAs	0.068	0.5	12.21	5.66	-6	-2

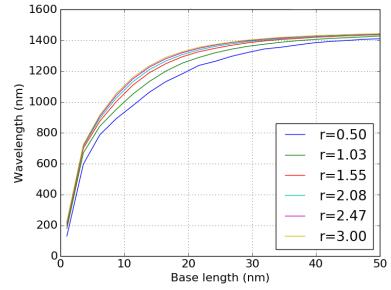


Fig. 5: Variation of emission wavelength with pyramidal quantum dot base size for different r values, without wetting layer

under the influence of strain is given by, $Eg = Eg_{unstrained} + \delta Eg$. Here $Eg_{unstrained}$ is the bandgap in unstrained condition. Different parameters used in the calculation are tabulated in table I.

D. Simulation Results

Fig. 5 shows the variation of emission wavelength with different pyramid base length parameterized for different r values, where r is the ratio of pyramid height to the base length. It can be noted that the general trend is the increase in wavelength with base length, but the increase seem to saturate beyond a certain base length (which depends on r). The corresponding results while wetting layer is shown in Fig. 6.

A curve fit was done for the same and the best match was found using the following equation but with different coefficients for base length above and below 10 nm.

$$\lambda(\text{nm}) = \frac{1240.8}{\alpha \exp(-\beta a) \exp(-\gamma r) \exp(-\zeta w_h) + \mu}$$

with coefficients $\alpha, \beta, \gamma, \zeta, \mu$ taking values as shown in table II. This is visualized in the Fig. 7 and Fig. 8 where Fig. 7 shows the fit for base length < 10 and Fig. 8 for that ≥ 10 .

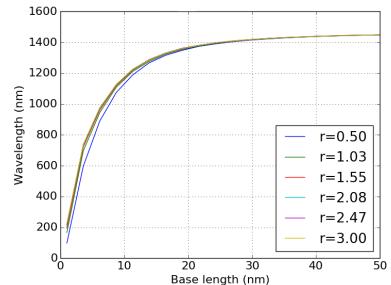


Fig. 6: Variation of emission wave length with base-length for different r values, with wetting layer.

TABLE II: Coefficients of curve fit

	α	β	γ	ζ	μ
Base size < 10 nm	18.00466793	0.77012327	0.2193027	0.41489452	1.16794573
Base size > 10 nm	1.65366554	0.12078621	0.0924974	1.78611513	0.85946292

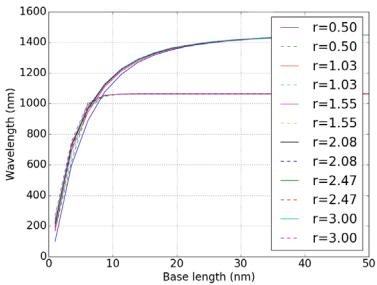


Fig. 7: Curve fit for results of wavelength vs base size, base length < 10 nm

Fig. 7 and Fig. 8 compares simulation and fit with different wh values for base length < 10 nm and ≥ 10 nm.

IV. CONCLUSION

Eigen modes simulation of time-independent Schrödinger wave equation in a InAs/GaAs quantum dot with pyramidal shape is simulated for different quantum dot sizes and found good emission wavelength match with experimental studies. Effect of wetting layer and lattice mismatch induced strain is also investigated. An empirical relationship is obtained relating pyramid's geometry to emission wavelength and which can give the geometry required for communication wavelength.

REFERENCES

- [1] Saima Beg, Syed Hasan Saeed, and MJ Siddiqui. Iii-v compound semiconductor laser heterostructures parametric performance evaluation for ingaas/gaas and algaaas/gaas. *Advances in Computational Sciences and Technology*, 10(10):2985–3013, 2017.
- [2] D Bhattacharyya, EA Avrutin, AC Bryce, JH Marsh, D Bimberg, F Heinrichsdorff, VM Ustinov, SV Zaitsev, NN Ledentsov, PS Kop'ev, et al. Spectral and dynamic properties of inas-gaas self-organized quantum-dot lasers. *IEEE Journal of selected topics in quantum electronics*, 5(3):648–657, 1999.
- [3] Nikolay Cherkashin, Shay Reboh, MJ Hýtch, Alain Claverie, VV Preobrazhenskii, MA Putyato, BR Semyagin, and VV Chaldyshev. Determination of stress, strain, and elemental distribution within in (ga) as quantum dots embedded in gaas using advanced transmission electron microscopy. *Applied Physics Letters*, 102(17):173115, 2013.

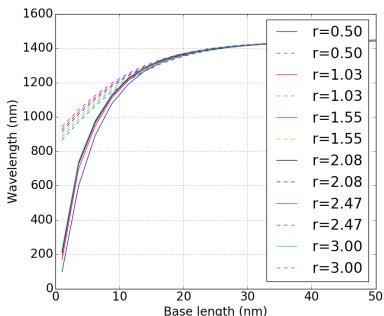


Fig. 8: Curve fit for results of wavelength vs base size, base length ≥ 10 nm

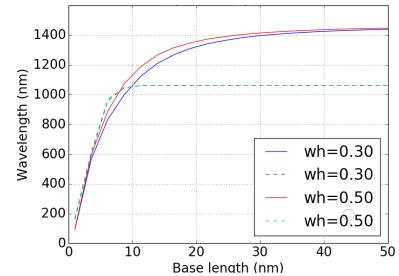


Fig. 9: Curve fit with different wetting later thickness, base length < 10 nm

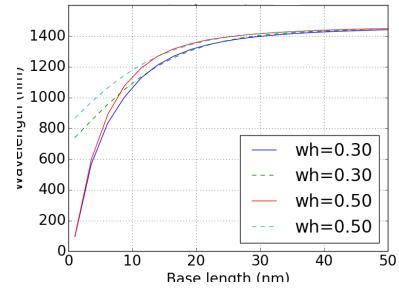


Fig. 10: Curve fit with different wetting later thickness, base length ≥ 10 nm

- [4] Matthew D Eisaman, JMAPS Fan, Alan Migdall, and Sergey V Polyakov. Invited review article: Single-photon sources and detectors. *Review of scientific instruments*, 82(7):071101, 2011.
- [5] Bouraoui Ilahi, Manel Souaf, Mourad Baira, Jawaher Alrashdi, Larbi Sfaxi, Abdulaziz Alhazaa, and Hassen Maaref. Evolution of inas/gaas qds size with the growth rate: a numerical investigation. *Journal of Nanomaterials*, 16(1):287, 2015.
- [6] SI Jung, HY Yeo, I Yun, JY Leem, IK Han, JS Kim, and JI Lee. Size distribution effects on self-assembled inas quantum dots. *Journal of Materials Science: Materials in Electronics*, 18(1):191–194, 2007.
- [7] Marco Rossetti, Lianhe Li, Alexander Markus, Andrea Fiore, Lorenzo Occhi, Christian Velez, Sergey Mikhlin, Igor Krestnikov, and Alexey Kovsh. Characterization and modeling of broad spectrum inas–gaas quantum-dot superluminescent diodes emitting at 1.2–1.3 μ m. *IEEE journal of quantum electronics*, 43(8):676–686, 2007.
- [8] Manel Souaf, Mourad Baira, Olfa Nasr, Mohamed Helmi Hadj Alouane, Hassen Maaref, Larbi Sfaxi, and Bouraoui Ilahi. Investigation of the inas/gaas quantum dot's size: dependence on the strain reducing layer's position. *Materials*, 8(8):4699–4709, 2015.
- [9] Introduction to quantum mechanics. Griffiths, David J and Schroeter, Darrell F, 2018. Cambridge University Press
- [10] Electronic structure and the properties of solids: the physics of the chemical bond, Harrison, Walter A, 2012. Courier Corporation
- [11] Theory and practice of finite elements, Ern, Alexandre and Guermond, Jean-Luc, 2013 Springer Science & Business Media
- [12] Semiconductor optoelectronic devices: introduction to physics and simulation, Piprek, Joachim, 2013 Elsevier
- [13] DJ Eaglesham and M Cerullo. Dislocation-free stranski-krastanow growth of ge on si (100). *Physical review letters*, 64(16):1943, 1990.
- [14] C-h Chiu, Z Huang, and CT Poh. Formation of nanostructures by the activated stranski-krastanow transition method. *Physical review letters*, 93(13):136105, 2004.
- [15] OE Shklyav, MJ Beck, M Asta, MJ Miksis, and PW Voorhees. Role of strain-dependent surface energies in ge/si (100) island formation. *Physical review letters*, 94(17):176102, 2005.
- [16] T Schwarz-Selinger, YL Foo, David G Cahill, and JE Greene. Surface mass transport and island nucleation during growth of ge on laser textured si (001). *Physical Review B*, 65(12):125317, 2002.

Adaptive Polarization Control for Coherent Optical Links with Polarization Multiplexed Carrier

Mehul Anghan

Electrical Engineering department

IIT Bombay

Mumbai, India

mpanghan28@gmail.com

Rashmi Kamran

Electrical Engineering department

IIT Bombay

Mumbai, India

rashmikamran@ee.iitb.ac.in

Nihar Gulati

Electrical Engineering department

IIT Kanpur

Kanpur, India

nihar@iitk.ac.in

Nandakumar Nambath

School of Electrical Sciences

IIT Goa

Ponda, India

npnandakumar@iitgoa.ac.in

Shalabh Gupta

Electrical Engineering department

IIT Bombay

Mumbai, India

shalabh@ee.iitb.ac.in

Abstract—Self-homodyne systems with polarization multiplexed carrier offer an local oscillator-less (LO-less) coherent receiver with simplified signal processing requirement that can be a good candidate for high-speed short-reach data center interconnects. The practical implementation of these systems is limited by the requirement of polarization control at the receiver end for separating the carrier and the modulated signal. In this paper, effect of polarization impairments in polarization diversity based systems is studied and modeled. A novel and practical adaptive polarization control technique based on optical power feedback from one polarization is proposed for polarization multiplexed carrier based systems and verified through simulation results. The application of the proposed concept is also experimentally demonstrated for a quadrature phase shift keying (QPSK) system with polarization multiplexed carrier.

I. INTRODUCTION

Merits of coherent modulation and demodulation techniques make them suitable for communication through optical fibres at high data rates [1]. Dual polarization quadrature phase shift keying (DP-QPSK) system has been commonly used for high data rates which utilizes diversity in both phase and polarization [2, 3]. This coherent technique uses a separate LO at the receiver. It also requires a carrier phase recovery and compensation module (CPRC) to overcome the effects of line-widths of the transmitter and receiver lasers and frequency offset between them. Need of the LO and the CPRC can be avoided in a polarization diversity based self-homodyne (SH) system, in which the carrier is polarization multiplexed with the modulated signal itself. Polarization impairments tied with the optical system and the channel cause mixing of the signals in two orthogonal polarizations that results in improper reception of the message symbols. Polarization demultiplexing techniques like constant modulus algorithm and decision directed algorithm in electrical domain can be used to compensate for these polarization impairments but it requires high speed signal processing [4, 5]. Circuit implementation of this high speed signal processing can be complex and

power hungry. A manual or automatic polarization controller device (for example EPC-400-11-1300/1550 from OZoptics) can be used to correct this kind of impairments. Some of these effects can be minimized by properly controlling the state of received polarization in the optical domain itself [6, 7]. Few polarization diversity based SH systems are demonstrated in prior works [8, 9] and have used manual polarization controller which is not a solution in a practical scenario. In another work [10], a polarization multiplexed carrier based SH system is demonstrated with direct detection receiver that faces issue of in-phase quadrature-phase (IQ) imbalance.

Adaptive polarization control is found to be very useful for SH coherent optical links with polarization multiplexed carrier. The proposed adaptive polarization control for a polarization diversity based SH system is presented in Fig. 1, in which the power of one of the polarizations (after converting to the electrical domain) is fed back to an electronically controlled polarization controller. In this paper, we have carried out modeling of the effects of polarization impairments in SH systems for short reach links. A technique based on minimization of the optical power received in one of the polarizations to control the state-of-polarization in short reach SH links is presented with the application of feedback based polarization control. A discrete time-gradient descent based algorithm is presented to achieve this minimization and validated using simulations. The proposed concept of polarization control has been experimentally verified for a polarization diversity based SH-QPSK system.

II. MODELLING OF POLARIZATION IMPAIRMENTS

Polarization impairments due to system components and fiber channel are discussed in this section. Polarization beam combiner (PBC) and polarization beam splitter (PBS) can mix the carrier and the modulated signal due to misalignment of reference axes as explained in Fig.2. The following equations

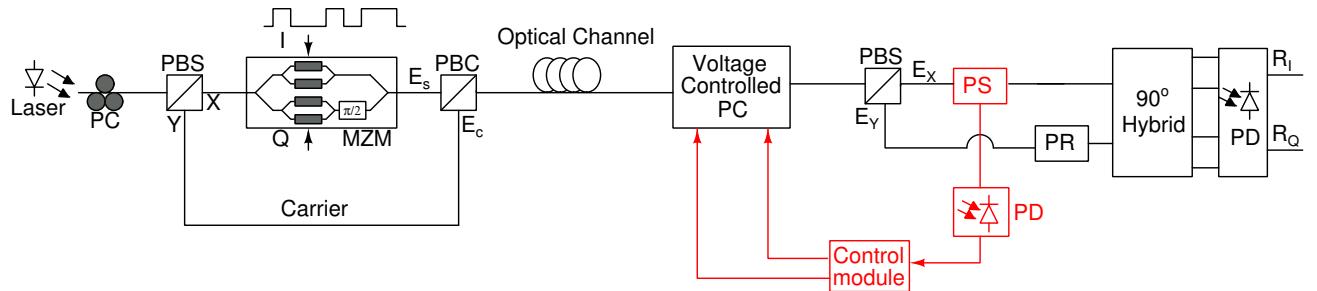


Fig. 1. A polarization diversity based SH-QPSK system with proposed adaptive polarization control. PC: polarization controller, PBS/PBC: polarization beam splitter/combiner, MZM: Mach-Zehnder modulator, PD: photo detector, PS: power splitter, and PR: polarization rotator.

represent the effect of PBS angle (θ) on the outputs of the PBS (PBS_x, PBS_y):

$$\begin{pmatrix} PBS_x \\ PBS_y \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix},$$

$$PBS_x = E_x \cos \theta - E_y \sin \theta,$$

$$PBS_y = E_x \sin \theta + E_y \cos \theta,$$

where E_x, E_y are the inputs to the PBS. Same phenomena can be explained for PBC also. A phase shift between two orthogonal polarizations due to fiber channel can also cause mixing. ϕ is the angle between the reference polarizations and the principle state of polarizations (PSPs). The overall effect due to the angles of PBS, PBC and ϕ can be represented as

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} e^{j\phi} & 0 \\ 0 & e^{-j\phi} \end{pmatrix} \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}.$$

Here the assumption is that device angles of PBS and PBC are the same. Outcome is the mixing of two polarizations.

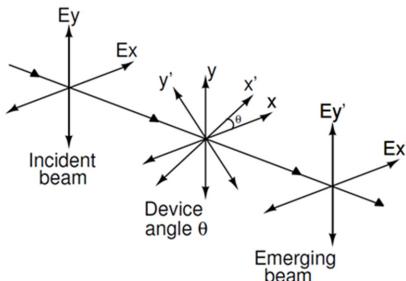


Fig. 2. Effect of the device angle θ on both polarizations. E_x and E_y are the inputs to the device with linear state of polarization. E_x' and E_y' are the outputs with changed state of polarization.

III. SEPARATION OF THE CARRIER AND THE MODULATED SIGNAL AT THE RECEIVER

This section discusses the technique for separating the carrier and the modulated signal for polarization diversity based SH systems. In this system as shown in Fig.1 the modulated signal is launched in one of the polarizations and the carrier signal is launched in the orthogonal polarization. Power of the launched carrier signal is higher than the power of the launched modulated signal. The power difference between

two polarizations is around 15 dB at the transmitter (due to modulator insertion loss). To separate the modulated signal and the carrier signal at the receiver power measurement in one of the polarizations can help. At the receiver, after polarization control, a PBS splits the signal into two polarization signals, which are applied to an optical hybrid. The optical hybrid couples these two input signals with 90° phase shift and outputs are converted to the electrical domain by balanced photo detectors for obtaining IQ data signals.

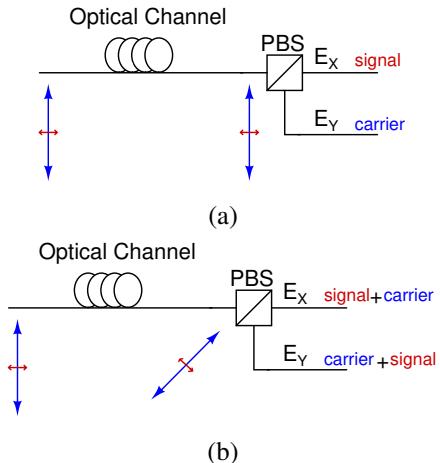


Fig. 3. Effect of the polarization control on separation of the modulated signal and the carrier at the receiver. (a) Ideally separated X and Y polarizations with linear state of polarization; and (b) mixed X and Y polarizations resulting in mixing of the carrier and the modulated signal with random state of polarization after the optical channel.

Ideal outputs of the PBS will be the low power modulated signal and the high power carrier as shown in Fig. 3. But, in practical scenario these two signals may get mixed resulting in a lower power difference between the two output branches of the PBS. Hence, to get the modulated signal back in one polarization a power minimization in that polarization at the receiver can be performed. The polarization diversity based SH-QPSK system is modeled in Simulink and VPITransmissionMaker™ with a three waveplate polarization controller (PC) module which has two control parameters for changing angles. Polarization impairments due to the system and the channel have been considered to be non-ideal. By varying two controls of the PC, optical power in one of the

outputs of the PBS has been measured and plotted as presented in Fig. 4.

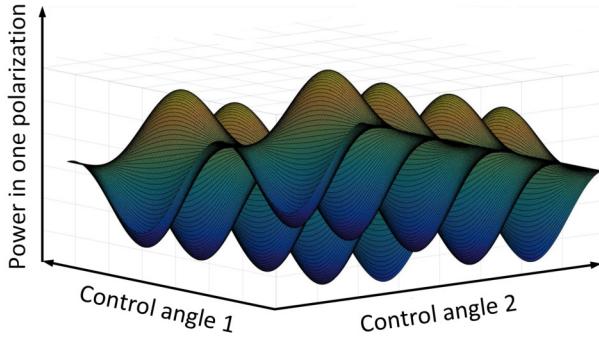


Fig. 4. Optical Power profile with respect to control angles of polarization controller at one of the output of the PBS. It can be seen that there are multiple minima of the same strength.

From Fig. 4 it can be seen that there are more than one minima with the same strength. At each minima equivalent matrix of the channel is converged to the following matrix.

$$\begin{pmatrix} e^{j\phi_1} & 0 \\ 0 & e^{-j\phi_1} \end{pmatrix}.$$

From this matrix we can see that after minimizing the optical power in one of the polarizations rotational effect and mixing of two polarization signals are removed and only phase shift in individual polarization remained. This phase shift can be removed using a CPRC module. Manual controlling of the PC is difficult to reach to any one of the minima. Polarization control with feedback has been implemented in VPItransmissionMaker™ to get the desired state of polarization. Optical power from any one of the polarizations is converted to the electrical domain using a photo-detector and based on the electrical signal it will adapt the control parameters of PC to achieve the minimum optical power.

IV. ADAPTIVE POLARIZATION CONTROL ALGORITHM

A flowchart of the discrete-time based gradient descent algorithm which has been used to find the minima and has been validated through simulations is shown in Fig. 5). Here C1 and C2 are control parameters of the PC, P is optical power in one polarization and μ is the step-size. The choice of the step-size value is very important for the algorithm. With a high value of the step-size the algorithm may converge to a wrong minimum. For a very small value of the step-size speed of polarization control will not be able to cope up with the change in state of polarization. State of polarization recovery time is dependent on the response times of electrically controlled PC and control circuitry for real time implementation.

V. SIMULATION RESULTS

Simulation has been performed for the 50 Gbaud SH-QPSK system with polarization multiplexed carrier shown in Fig. 1 for 20 km distance in VPItransmissionMaker™ with all non-idealities ON and added noise (OSNR 25 dB). A laser power

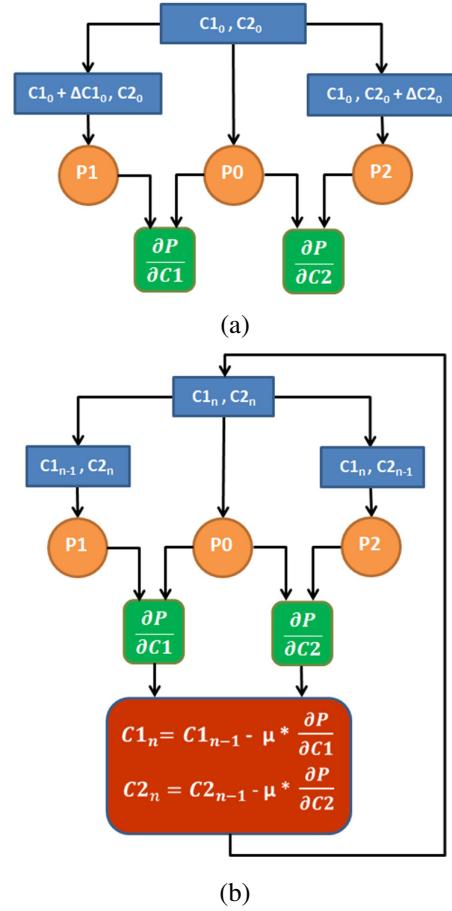


Fig. 5. Flowchart of the gradient descent algorithm for minimizing power in one of the polarizations. (a) Initialization of coefficients based on the gradient; and (b) adaptation of coefficient values to achieve the minimum power according to the input state of polarization. C1 and C2 are control parameters of the PC, P is optical power in one polarization and μ is the step-size.

of 10 mW and single mode fiber (SMF) with a dispersion coefficient of 16 ps/km.nm and attenuation of 0.2 dB/km, were used for simulations. Polarization mode dispersion and non-linearity effects were kept ON for the simulations. The adaptive control algorithm has been implemented by using the script editor feature of VPItransmissionMaker™. Fig. 7 shows the spectrum of both the polarizations without PC which indicates the mixing of the modulated signal and the carrier. A proper separation of the modulated signal and the carrier is clearly visible in Fig. 8 by using the proposed adaptive polarization control.

The chromatic dispersion (CD) increases with data rate and fiber length. CD does not cause a mixing of the carrier and the modulated signal. The dispersion effect can also be observed in the shape of the spectrum as the high data rate with 20 km distance has been considered in simulations. It can be concluded from the results that the minimization of power in one polarization is able to separate the carrier and the modulated signal even if there is a significant amount of chromatic dispersion.

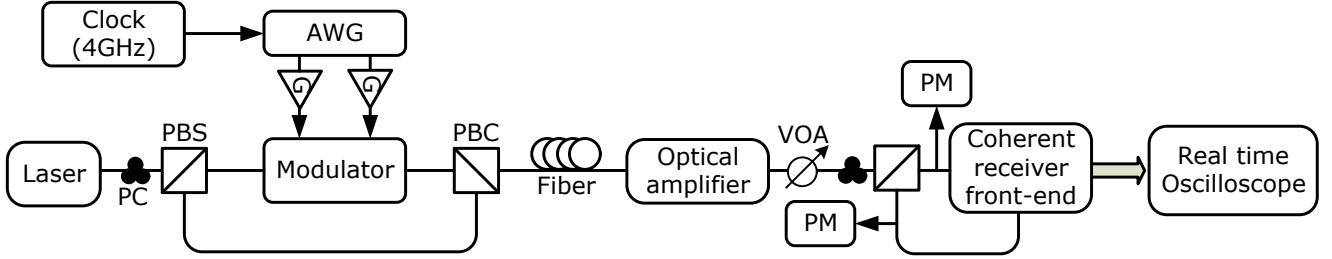


Fig. 6. Experimental setup for an SH-QPSK system with polarization multiplexed carrier. AWG: arbitrary waveform generator, PC: polarization controller, PBS/PBC: polarization beam combiner/splitter, VOA: variable optical attenuator, and PM: power meter.

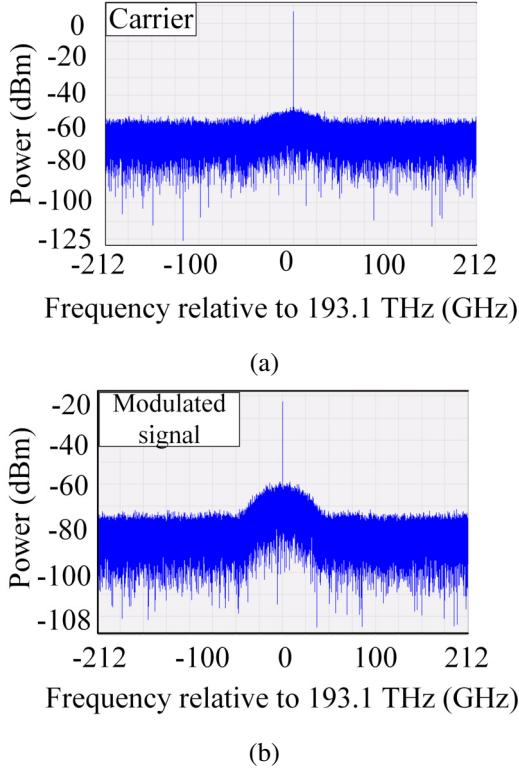


Fig. 7. Spectrum of X and Y polarizations for an SH-QPSK system without polarization control for data rate: 100 Gbps, distance: 20 km and OSNR: 25 dB. (a) LO mixed with the modulated signal; and (b) the modulated signal mixed with the LO.

VI. EXPERIMENTAL RESULTS

The concept has also been verified by an experimental demonstration of the polarization diversity based SH-QPSK system shown in Fig. 6. In this setup a manual three-waveplate PC has been used to minimize the power in one of the polarizations. The output from an external cavity laser (ECL) SFL1550P having power of 13.02 dBm is split into two polarization using a PBS. Output power of the PBS branches are set equal for attaining linear state of polarization using a PC connected to the laser source. One of the branches of the PBS is connected to a QPSK modulator LN86S-FC which is driven by two amplified RF signals of 4 Gb/s data rate generated by using an arbitrary wave form generator (Euvis AWG 801). Another branch of the PBS output is directly combined with

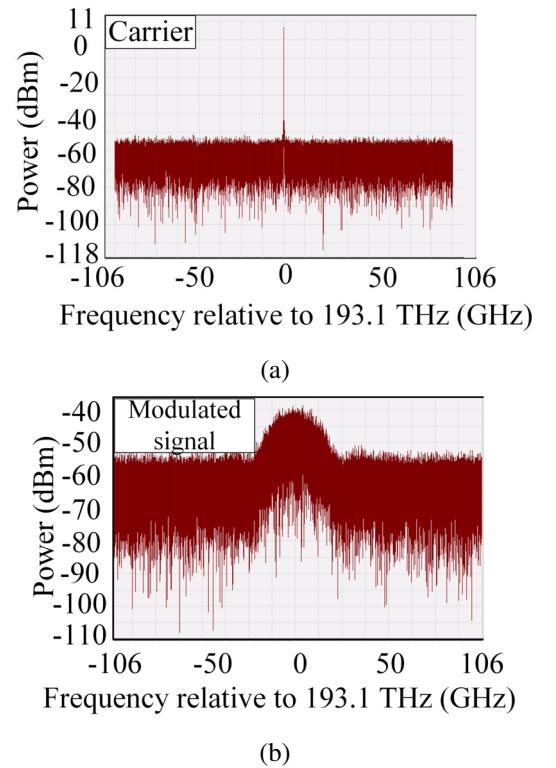


Fig. 8. Spectrum of X and Y polarizations for an SH-QPSK system after minimizing power in one polarization for data rate: 100 Gbps, distance: 20 km and OSNR: 25 dB. (a) Separated LO; and (b) separated modulated signal.

the output of the modulator using a PBC. The received signal from the single mode fiber is split into orthogonal polarizations (X and Y) using a PBS after manually controlling the PC. One of the PBS outputs having lower power is connected to the signal port of the receiver front-end CPRV1222A and the PBS output having a higher power is connected to the LO port of the receiver front-end. Received signals are captured by a real-time oscilloscope with and without polarization control.

Manually rotatable three-paddles based polarization controller is connected to the fiber output to separate the carrier and the modulated data. To fulfill this requirement two power meters are connected to the output branches of the PBS at the receiver side as shown in Fig. 6. Now by rotating the paddles of the PC power in one polarization is minimized

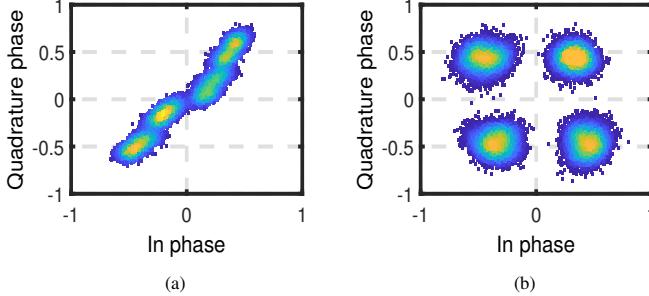


Fig. 9. Performance of a 4 Gbps SH-QPSK system for 30 km distance. Constellation diagrams: (a) received signal without minimizing power (power difference: 2.64 dB); and (b) received signal after power minimizing (power difference: 13.2 dB, error vector magnitude (EVM) after phase correction: 26.66%).

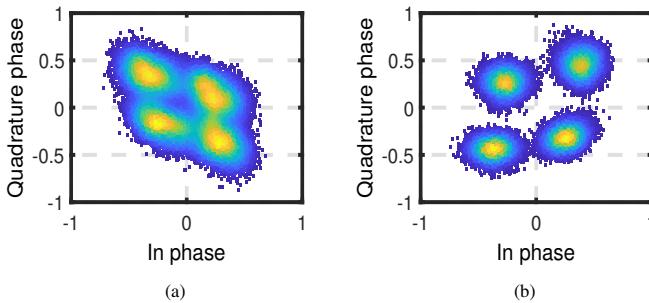


Fig. 10. Performance of an 8 Gbps SH-QPSK system for 30 km distance. Constellation diagrams: (a) received signal without minimizing power (power difference: 5.32 dB); and (b) received signal after power minimizing (power difference: 13.03 dB, EVM after phase correction: 34.16%).

and power in the other polarization is maximized with a resulting power difference of around 13 dB. Experiment has been performed for 30 km SMF with 4 Gbps and 8 Gbps SH-QPSK systems due to setup limitations. Results are presented in Fig.9 and Fig.10. Overall performance of the setup can be improved with the use of single polarization receiver front end. Difference and improvements in the constellations is clearly observed in results with and without power minimization in one polarization. Real time implementation of an SH-system with proposed adaptive polarization control technique requires a voltage controlled PC with low response time and high speed control circuitry.

VII. CONCLUSION

The proposed adaptive polarization control technique for polarization diversity based SH systems is practically feasible for implementation in real-time systems. The method has been successfully validated for an SH-QPSK system with polarization multiplexed carrier. This work opens up a choice to employ SH systems for low power high capacity data center interconnects.

REFERENCES

- [1] K. Kikuchi, "Fundamentals of Coherent Optical Fiber Communications," *J. Lightwave Technol.*, vol. 34, no. 1, pp. 157–179, Jan 2016.
- [2] Takahashi *et al.*, "Compact 100-Gb/s DP-QPSK intradyne coherent receiver module employing Si waveguide," pp. 1–3, 09 ECOC 2015.
- [3] N. Nambath *et al.*, "Analog Domain Signal Processing-Based Low-Power 100-Gb/s DP-QPSK Receiver," *J. Lightwave Technol.*, vol. 33, no. 15, pp. 3189–3197, Aug 2015.
- [4] Noe *et al.*, "Electronic polarization control algorithms for coherent optical transmission," vol. 16, pp. 1193 – 1200, 11 2010.
- [5] R. S. Lus *et al.*, "Digital Self-Homodyne Detection," *IEEE Photonics Technology Letters*, vol. 27, no. 6, pp. 608–611, 2015.
- [6] M. Yagi *et al.*, "Field Trial of 160-Gbit/s, Polarization-Division Multiplexed RZ-DQPSK Transmission System using Automatic Polarization Control," in *OFC/NFOEC 2008 - 2008 Conference on Optical Fiber Communication/National Fiber Optic Engineers Conference*, Feb 2008, pp. 1–3.
- [7] B. Koch *et al.*, "Endless Optical Polarization Control at 56 krad/s, Over 50 Gigadian, and Demultiplex of 112-Gb/s PDM-RZ-DQPSK Signals at 3.5 krad/s," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 16, no. 5, pp. 1158–1163, Sept 2010.
- [8] T. Miyazaki, "Linewidth-tolerant QPSK homodyne transmission using a polarization-multiplexed pilot carrier," *IEEE Photonics Technology Letters*, vol. 18, no. 2, pp. 388–390, Jan 2006.
- [9] M. Nakamura *et al.*, "Experimental demonstration of 16-QAM transmission with a single dual-drive Mach-Zehnder modulator," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, March 2011, pp. 1–3.
- [10] P. Boffi *et al.*, "Self-homodyne coherent system based on polarization-multiplexed pilot-carrier," in *Fotonica AEIT Italian Conference on Photonics Technologies*, 2015, May 2015, pp. 1–3.

Analysis of Beam Wander Effect of Flat-topped Multi-Gaussian Beam for FSO Communication Link

Arka Mukherjee

Bharti School of Telecom Tech. & Mgmt.
Indian Institute of Technology Delhi
New Delhi, India
bsy157547@dbst.iitd.ac.in

Subrat Kar

Dept. of Electrical Engg.
Indian Institute of Technology Delhi
New Delhi, India
subrat@ee.iitd.ac.in

Virander Kumar Jain

Dept. of Electrical Engg.
Indian Institute of Technology Delhi
New Delhi, India
vkjain@ee.iitd.ac.in

Abstract—Atmospheric turbulence causes severe impairment of FSO communication link. By using the earlier model for the Gaussian beam, we analyze the beam wander effect for the flat-topped multi-Gaussian beam. The link availability decreases drastically in high turbulence regime for a Gaussian beam due to turbulence induced beam wander. In this paper, we model each turbulent eddy as a thin dielectric lens with Gaussian shaped refractive index profile and assume there are several sheets of eddies throughout the propagation path. We consider uniformly distributed eddy positions in a laminar sheet with Chi-Square distributed eddy sizes. We graphically demonstrate beam wander characteristics for different beam sizes and orders of the flat-topped multi-Gaussian beam in all three turbulence regimes characterized by different refractive index structure parameter values. Our results show that the flat-topped beam has a limited advantage in weak and moderate turbulence regimes. But it has a significant advantage in high turbulence regime to mitigate link outage due to beam wander.

Index Terms—Atmospheric turbulence, beam wander, free space optical communication

I. INTRODUCTION

There have been panoptic studies on the effect of various atmospheric phenomena on laser beam propagation through free space for years [1]–[4]. Free space optical (FSO) communication has evolved as a robust alternative to radio frequency (RF) communication due to many advantages like higher power efficiency, more bandwidth, better security because of very low beam divergence, and unlicensed spectrum [5]–[7]. However, different atmospheric effects like absorption, scattering, and turbulence cause severe performance degradation in FSO link. Among these effects, atmospheric turbulence has the most impact and is associated with effects like beam spreading, beam wander, and beam scintillation.

The earth's surface gets heated by absorbing solar radiations, and the adjacent layer of air gets warm. Due to convection, this warmer air rises above and mixes turbulently with relatively cooler surrounding air causing random fluctuations of temperature. Locally unstable air volumes or inhomogeneities resulting from this turbulence are considered as “eddies” of different sizes, temperatures, and indices of refraction [8]. Due to inertial forces, larger eddies break into smaller eddies and form a continuum of eddy

sizes between a macroscale (or the outer scale of turbulence, L_0) and a microscale (or the inner scale of the turbulence, l_0). Eddies bounded by these two limits are considered statistically homogeneous and isotropic [9]. When eddies are much smaller than the beam size, diffraction effect dominates, and the beam spreads. If eddies are much larger than the beam size, it will deflect the whole beam randomly from its original path, and the beam might miss the receiver due to beam wander. If the eddy size is of the same order of beam size, then the eddy acts like a lens that continuously focuses and de-focuses the incoming beam. It causes random irradiance fluctuations at the receiver causing beam scintillation [10]. The scintillation effect has been studied extensively using probability density functions (PDFs) like log-normal, negative-exponential, Gamma-Gamma, I-K, and Weibull distributions [11, 12]. Malaga (\mathcal{M}) distribution [13] and Double Generalized Gamma distribution [14] are new additions in this list.

Many researchers have studied the effect of beam wander in FSO communication link both theoretically and experimentally. In [15], geometrical optics has been used to estimate beam wander in weak turbulence regime. Beam wander characteristics of dark hollow, flat-topped, and annular beams have been studied using first and second order statistical moments [16]. The same method has been applied for cos and cosh-Gaussian beams [17]. Another study has demonstrated beam wander analysis for J_0 - and I_0 -Bessel Gaussian beam using the same method [18]. An analytical-numerical hybrid technique has been used for beam wander calculation for a satellite uplink [19]. Experimental studies of beam wander under different turbulence conditions are also conducted [20, 21]. A beam wander model using extended Rytov theory with small-scale and large-scale filters is presented recently [22]. In [23], the authors report a beam wander model for Gaussian beam using the first principle. The existing model [23] indicates that the link availability decreases drastically in high turbulence regime due to beam wander effect.

In this paper, we model each eddy as a dielectric lens with Gaussian shaped refractive index (RI) profile and characterize the beam wander effect for the flat-topped multi-Gaussian beam (FMGB) of different orders and beam widths in all

three turbulence regimes characterized by different refractive index structure parameter (C_n^2) values. Our study shows that the advantage of the FMGB is limited up to certain beam width in weak and moderate turbulence regimes. However, the FMGB has a significant advantage in high turbulence regime to mitigate link outage due to beam wander.

II. THEORY

A. Classification of Atmospheric Turbulence

RI of the propagation path is one of the most significant parameters for the study of laser beam propagation in the atmosphere. Due to the presence of turbulence (and hence random fluctuations of the temperature), RI of the medium varies randomly. For optical and IR wavelengths, the dependence of RI (n) at a point r in space with pressure and temperature is [9]

$$n(r) = 1 + 79 \times 10^{-6} \frac{P(r)}{T(r)} \quad (1)$$

where P is the pressure in millibars, and T is the temperature in Kelvin. As the variation of pressure is negligibly small, the temperature variation becomes the key reason for the change in RI.

Depending on the values of C_n^2 ($\text{m}^{-2/3}$), the turbulence is broadly classified into following three regimes [24]:

$$\begin{aligned} C_n^2 \geq 10^{-13} &\rightarrow \text{Strong turbulence} \\ 10^{-16} < C_n^2 < 10^{-13} &\rightarrow \text{Moderate turbulence} \\ C_n^2 \leq 10^{-16} &\rightarrow \text{Weak turbulence} \end{aligned}$$

B. Flat-topped multi-Gaussian Beam

The multi-Gaussian function is a finite sum of Gaussian components side by side with the same width, phase curvature, and absolute phase. The general expression of a two-dimensional multi-Gaussian function is [25]

$$MG(x, y) = \frac{\sum_{p=-N_1}^{N_1} \sum_{q=-M_1}^{M_1} \exp \left[-\frac{(x-pw)^2 + (y-qw)^2}{w^2} \right]}{\sum_{p=-N_1}^{N_1} \sum_{q=-M_1}^{M_1} \exp \left[-(p^2 + q^2) \right]} \quad (2)$$

where N_1, M_1 are the order of the multi-Gaussian function in the x -direction and the y -direction, respectively, w is the width of the individual Gaussian component, and p and q represent the offset of the corresponding Gaussian component. When $M_1 = N_1$, the shape becomes square. The composite width of the multi-Gaussian profile W is related to w by

$$W = \frac{w}{N_1 + \left\{ 1 - \ln \left[\sum_{p=-N_1}^{N_1} \exp(-p^2) \right] \right\}^{1/2}} \quad (3)$$

In Fig. 1, multi-Gaussian profiles of order 2, 5, and 9 are shown. It is evident from the figure that as the order N_1 increases, the edge becomes sharper.

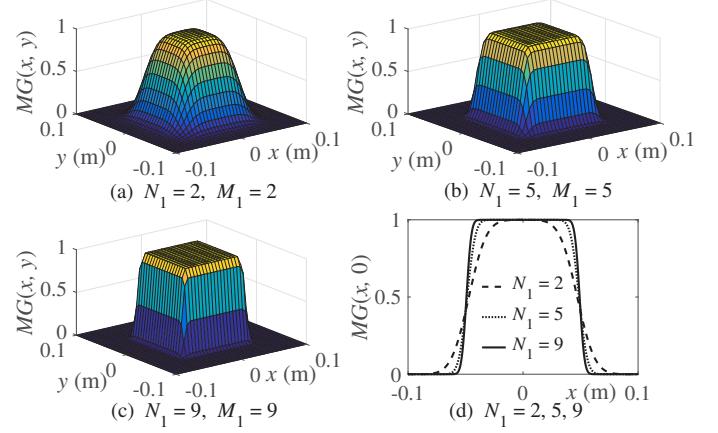


Fig. 1: Multi-Gaussian shapes of different orders when $N_1 = M_1$ and $W = 5$ cm, (a), (b), and (c) show the three dimensional shapes, and (d) is the profile of (a), (b), and (c)

C. Modeling of Eddies

An eddy is typically modeled as a thin dielectric lens with Gaussian shaped RI profile. The radius R of the lens is defined to be the $1/e$ distance from the centre of index distribution, and the focal length, f of the lens is given by [26]

$$f = R/2\delta n \quad (4)$$

where $\delta n = n_1 - n_0$, n_1 is the peak RI associated with the eddy, and n_0 is the RI of free space which is unity. In general, $\delta n \leq 10^{-6}$, and the resultant focal lengths of eddies are of the order of few thousands of meters. We assume eddies are distributed over thin sheets perpendicular to the direction of propagation. For a horizontal terrestrial link of length L , there are on an average $L/2R$ numbers of such eddies encountered by the propagating beam. Fig. 2 depicts the scenario for one sheet of eddies that act as thin dielectric lenses with a Gaussian shaped RI profile.

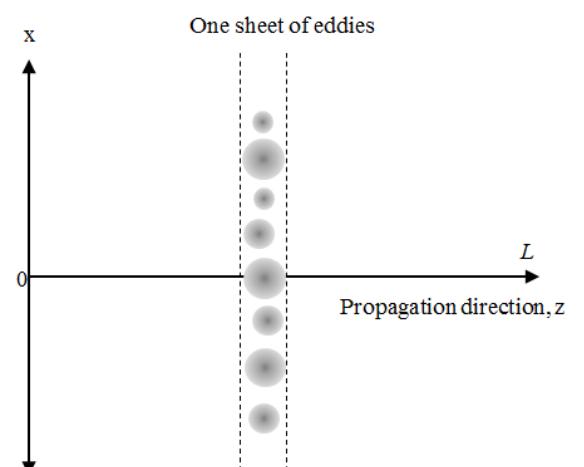


Fig. 2: One sheet of eddies that act on the propagating beam as thin lenses with Gaussian shaped RI profile

D. Distribution of Eddies

The eddy sizes are considered to be Chi-Square distributed [23, 27]. δn is assumed to be negative exponentially distributed [23].

For a random variable R , the PDF of Chi-Square distribution is given by

$$f_R(R) = \frac{R^{(m-2)/2} \exp(-R/2)}{2^{m/2} \Gamma(m/2)} \quad (5)$$

where m is the degree of freedom parameter and taken to be $1/3 \times ((C_n^2)^{-2/3} \times \lambda)^{1/2}$. $\Gamma(\cdot)$ denotes the Gamma function. For a random variable δn , the PDF of negative exponential distribution is given by

$$f_{\delta n}(\delta n) = \frac{1}{\beta} \exp(-\frac{\delta n}{\beta}) \quad (6)$$

where β is the mean of the PDF and taken to be $(C_n^2/1000)^{1/3}$.

To simulate the beam wander effect under worst-case scenario when a propagating beam encounters a maximum number of turbulent eddies for a given propagation distance and turbulence condition, we assume the centres of eddies follow a uniform distribution within a circle of radius R centred at the propagation axis.

E. Analytical background

In free space, a laser beam has divergence approximated by λ/D where λ is the wavelength of the laser beam, and D is the transmit aperture diameter. When turbulence is present, eddies larger than the beam diameter tend to deflect the beam whereas smaller eddies tend to spread the beam [28]. If the exposure time is much lower than the time, δt where $\delta t = D/|v|$, and v is transverse wind velocity, the beam spot wanders in the receiver aperture plane with beam shape being almost same as would be in free space. If the exposure time is much larger than δt , we would see a single beam spot with a larger diameter as shown in Fig. 3 [29].

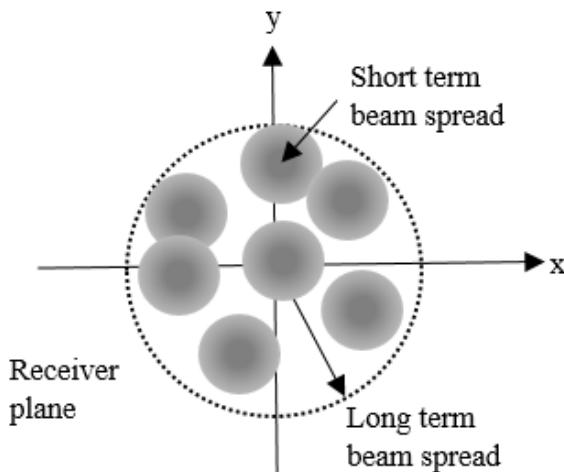


Fig. 3: Beam wander effect

The mean square beam wander, $\langle b_w^2 \rangle$ at an axial position at $z = L$, is related to on axis first and second order statistical moments, $E_1(0, 0)$ and $E_2(0, 0)$, respectively, and it is given by [9]

$$\langle b_w^2 \rangle = [-2E_1(0, 0) - E_2(0, 0)] \alpha_r^2(L) \quad (7)$$

where $\alpha_r(L)$ is the beam size at $z = L$. The derivations leading to $E_1(0, 0)$ and $E_2(0, 0)$ are quite complicated and lengthy. For a fundamental Gaussian beam, (7) can be simplified as [16]

$$\langle b_w^2 \rangle = 2.42 C_n^2 L^3 \alpha_s^{-1/3} \quad (8)$$

where α_s is the transmitted Gaussian beam size.

III. SIMULATION RESULTS

We have used Matlab for simulating the beam wander effect at a wavelength of 1550 nm for 1000 pulses for a horizontal terrestrial link length of 2 km. Commonly used aperture diameters [30, 31] have been taken to calculate the link availability which is defined as

$$\text{Link availability} = \frac{\text{No of pulses captured by aperture}}{\text{Total no of pulses (N)}} \times 100\% \quad (9)$$

For the horizontal link, the value of C_n^2 is considered to be constant. We use $ABCD$ matrix method [32] to calculate the effect of eddy lenses on the propagating beam. In Figs. 4-6, we have shown beam wander characteristics of the FMGB of order 5 and 9 along with the fundamental Gaussian beam in weak, moderate, and high turbulence regimes, respectively. It is evident from Figs. 4 and 5 that the use of the FMGB is advantageous for transmitted beam size less than 5 mm and 1 cm, respectively, where the RMS beam wander of the fundamental Gaussian beam is more than that of the FMGB.

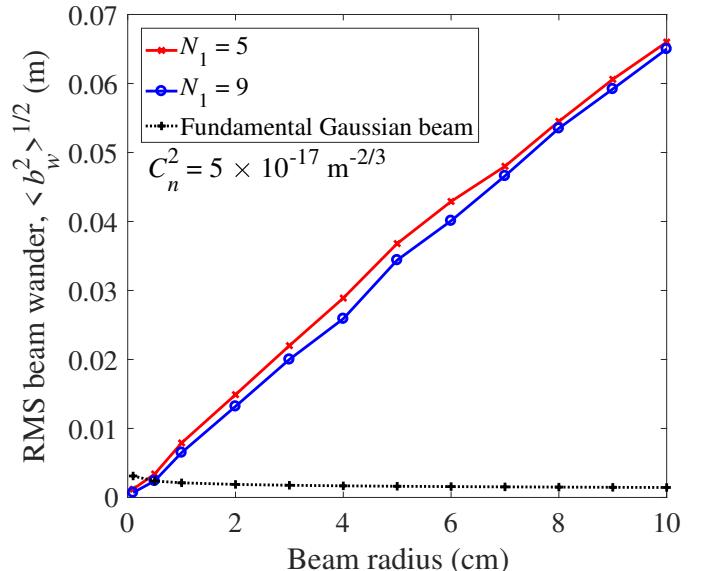


Fig. 4: RMS beam wander for FMGB of order 5 and 9 and fundamental Gaussian beam in weak turbulence

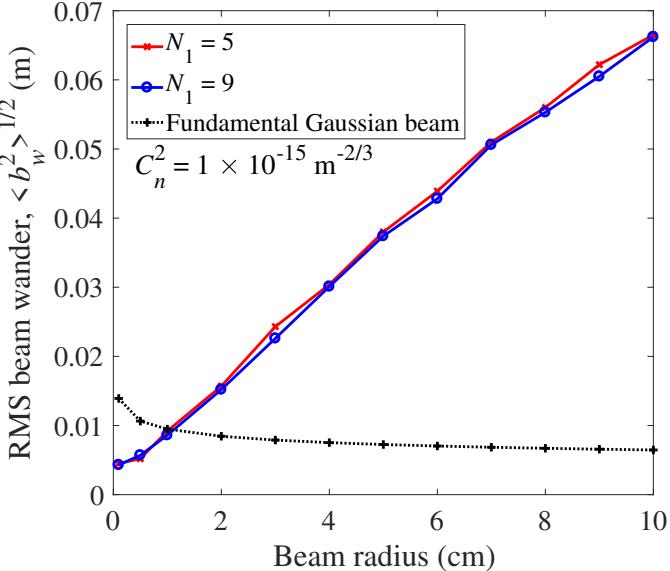


Fig. 5: RMS beam wander for FMGB of order 5 and 9 and fundamental Gaussian beam in moderate turbulence

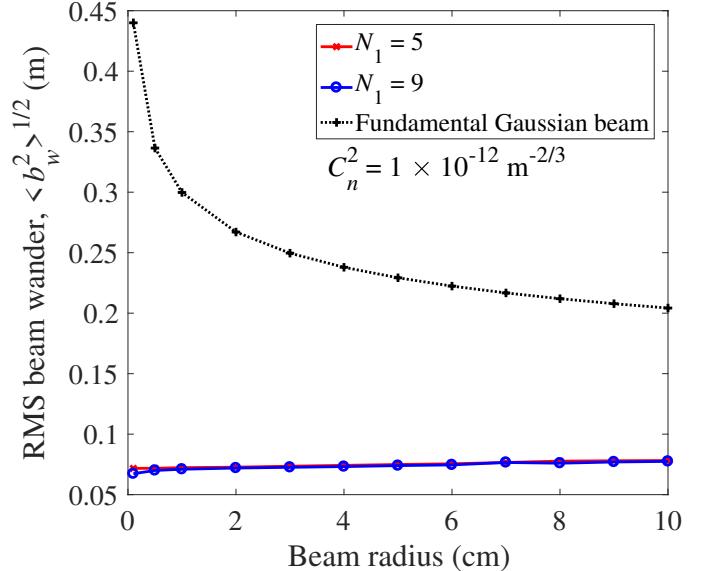


Fig. 6: RMS beam wander for the FMGB of order 5 and 9 and fundamental Gaussian beam in high turbulence

In other words, as the beam size increases the advantage of using the FMGB gradually diminishes in weak and moderate turbulence regimes. Whereas, in high turbulence regime, the FMGB has significantly low RMS beam wander than the fundamental Gaussian beam irrespective of the beam size. The FMGB of order 9 faces relatively lower beam wander than the FMGB of order 5. However, choosing a beam of order higher than 9 has a minimal effect on the beam wander performance than the cost associated with it. Link availability in high turbulence regime ($C_n^2 = 10^{-12} \text{ m}^{-2/3}$) for $W = 1$ cm and $N_1 = 9$ is 40%, 57%, and 90% for a receiver aperture diameter of 70 mm, 140 mm, and 280 mm, respectively. The corresponding values for the fundamental Gaussian beam are 9.1%, 30.3%, and 73.6% [23]. Thus, the FMGB of proper order and beam size has a significant role to mitigate beam wander effect due to atmospheric turbulence. Additionally, it helps in minimizing background noise incorporated in the system due to the use of larger aperture for a required link availability.

IV. CONCLUSION

Beam wander characteristics of the FMGB were studied in all three turbulence regimes by modeling the eddy sizes and refractive index profiles. It saved the tedious calculation of statistical moments in beam wander analysis. Our simulation results agree with the outcomes of previous works on different types of flat-topped beams. As turbulence strength increases, the link availability for an FSO link reduces abruptly for the fundamental Gaussian beam. By transmitting the FMGB, the link availability can be increased even in high turbulence. However, in weak and moderate turbulence regimes, the benefit of using the FMGB is limited up to small transmitted

beam size. So, by carefully selecting the transmitted beam shape, order, and size, we can achieve considerably higher link availability for an FSO link in all three turbulence regimes. The use of the FMGB also minimizes the effect of background noise incorporated in the FSO communication system when using aperture averaging, i.e., larger aperture area to achieve higher link availability.

REFERENCES

- [1] V. W. S. Chan, "Free-space optical communications," *J. Lightw. Technol.*, vol. 24, no. 12, pp. 4750–4762, Dec. 2006.
- [2] H. Henniger and O. Wilfert, "An introduction to free-space optical communications," *Radioengineering*, vol. 19, no. 2, pp. 203–212, June 2010.
- [3] H. E. Nistazakis, T. A. Tsiftsis, and G. S. Tombras, "Performance analysis of free-space optical communication systems over atmospheric turbulence channels," *IET Commun.*, vol. 3, no. 8, pp. 1402–1409, Aug. 2009.
- [4] A. K. Majumdar, "Free-space laser communication performance in the atmospheric channel," *J. Opt. Fiber Commun. Rep.*, vol. 2, no. 4, pp. 345–396, Oct. 2005.
- [5] H. Kaushal, G. Kaddoum, V. K. Jain, and S. Kar, "Experimental investigation of optimum beam size for FSO uplink," *Opt. Commun.*, vol. 400, pp. 106–114, Oct. 2017.
- [6] A. G. Alkholidi and K. S. Altowij, "Free space optical communications—theory and practices," in *Contemporary Issues in Wireless Communication*, InTech, 2014, pp. 159–212.
- [7] I. E. Lee, Z. Ghassemlooy, W. Pang, M. Khalighi, and S. K. Liaw, "Effects of aperture averaging and beam width on a partially coherent Gaussian beam over free-space optical links with turbulence and pointing errors," *Appl. Opt.*, vol. 55, no. 1, pp. 1–9, Jan. 2016.
- [8] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, "Channel modelling," in *Optical Wireless Communications: System and Channel Modelling with Matlab®*, Boca Raton, FL: CRC Press, 2012, pp. 77–159.
- [9] L. C. Andrews and R. L. Phillips, *Laser Beam Propagation through Random Media*, Bellingham, WA: SPIE Press, 2005.
- [10] H. Kaushal, V. K. Jain, and S. Kar, "Free-space optical channel models," in *Free Space Optical Communication*, B. Mukherjee, Ed. India: Springer, 2017, pp. 41–89.

- [11] M. A. Al-Habash, L. C. Andrews, and R. L. Phillips, "Mathematical model for the irradiance probability density function of a laser beam propagating through turbulent media," *Opt. Eng.*, vol. 40, no. 8, pp. 1554–1562, Aug. 2001.
- [12] R. Barrios, "Exponentiated Weibull fading channel model in free-space optical communications under atmospheric turbulence," Ph.D. dissertation, Polytechnic Univ. of Catalonia, Spain, 2013.
- [13] A. Jurado-Nava, J. M. Garrido-Balsells, M. Castillo-Vázquez, A. Puerta-Notario, I. T. Monroy, and J. J. V. Olmos, "Fade statistics of M -turbulent optical links," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 1, June 2017.
- [14] M. A. Kashani, M. Uysal, and M. Kavehrad, "A novel statistical channel model for turbulence-induced fading in free-space optical systems," *J. Lightw. Technol.*, vol. 33, no. 11, pp. 2303–2312, June 2015.
- [15] J. H. Churnside and R. J. Latitis, "Wander of an optical beam in the turbulent atmosphere", *Appl. Opt.*, vol. 29, no. 7, pp. 926–930, Mar. 1990.
- [16] H. T. Eyyuboğlu and C. Z. Çil, "Beam wander of dark hollow, flat-topped and annular beams," *Appl. Phys. B*, vol. 93, pp. 595–604, Nov. 2008.
- [17] C. Z. Çil, H. T. Eyyuboğlu, Y. Baykal, and Y. Cai, "Beam wander characteristics of cos and cosh-Gaussian beams," *Appl. Phys. B*, vol. 95, no. 4, pp. 763–771, June 2009.
- [18] C. Z. Çil, H. T. Eyyuboğlu, Y. Baykal, O. Korotkova, and Y. Cai, "Beam wander of J_0 - and I_0 -Bessel Gaussian beams propagating in turbulent atmosphere," *Appl. Phys. B*, vol. 98, no. 1, pp. 195–202, Jan. 2010.
- [19] F. Dios, J. A. Rubio, A. Rodríguez, and A. Comerón, "Scintillation and beam-wander analysis in an optical ground station-satellite uplink," *Appl. Opt.*, vol. 43, no. 19, pp. 3866–3873, July 2004.
- [20] H. Kaushal, V. Kumar, A. Dutta, H. Aennam, V. K. Jain, S. Kar, and J. Joseph, "Experimental study on beam wander under varying atmospheric turbulence conditions," *IEEE Photon. Technol. Lett.*, vol. 23, no. 22, pp. 1691–1693, Nov. 2011.
- [21] X. Liu, F. Wang, C. Wei, and Y. Cai, "Experimental study of turbulence-induced beam wander and deformation of a partially coherent beam," *Opt. Lett.*, vol. 39, no. 11, pp. 3336–3339, June 2014.
- [22] Q. Wang, Y. Zhu, and Y. Zhang, "Precision wander model of beam wave in the weak to strong turbulence of atmosphere," in *Opt. Lett.*, vol. 42, no. 16, pp. 3213–3216, Aug. 2017.
- [23] A. Mukherjee, S. Kar, and V. K. Jain, "Analysis of FSO link under atmospheric turbulence from first principle," in *Proc. OPJ-OSA Joint Symposia on Nanophotonics and Digital Photonics*, Japan, 2017, paper 30aOD6.
- [24] D. H. Titterton, "Laser beam propagation," in *Military Laser Technology and Systems*. Norwood, MA: Artech House, 2015, pp. 165–196.
- [25] Y. Gao, B. Zhu, D. Liu, and Z. Lin, "Propagation of flat-topped multi-Gaussian beams through a double-lens system with apertures," *Opt. Exp.*, vol. 17, no. 15, pp. 12753–12766, July 2009.
- [26] L. C. Andrews, R. L. Phillips, and C. Y. Hopen, "Modeling optical scintillation," in *Laser Beam Scintillation with Applications*, Bellingham, WA: SPIE Press, 2001, vol. PM99, pp. 67–96.
- [27] M. Firoozmand and M. N. Moghadasi, "Modeling and simulation of fading due of atmospheric turbulence by Chi-Squared and exponential pdf for a FSO link," *NNGT Int. J. Netw. Comput.*, vol. 2, Feb. 2015.
- [28] P. Latsa Babu and B. Srinivasan, "Characterizing the atmospheric effects on laser beam propagation for free space optical communication," in *National Conf. Commun. (NCC)*, India, Feb. 2008, pp. 332–334.
- [29] R. L. Fante, "Electromagnetic beam propagation in turbulent media," *Proc. of the IEEE*, vol. 63, no. 12, pp. 1669–1692, Dec. 1975.
- [30] R. Srinivasan and D. Sridharan, "The climate effects on line of sight (LOS) in FSO communication," in *IEEE Int. Conf. Computational Intell. Comput. Res.*, India, 2010.
- [31] A. Prokeš, "Terrestrial free space optics architecture," A thesis of a talk, Brno Univ. of Technology, Czech Republic, 2009.
- [32] E. Hecht, *Ray tracing and ABCD matrix* [online]. Available: www.ece.tamu.edu/~prhemmer/elen489_501/17-abcd.ppt. [Accessed: Sep. 13, 2018].

Qubits through Queues: The Capacity of Channels with Waiting Time Dependent Errors

Krishna Jagannathan, Avhishek Chatterjee
 Department of Electrical Engineering
 IIT Madras, Chennai, India

Prabha Mandayam
 Department of Physics
 IIT Madras, Chennai, India

Abstract—We consider a setting where qubits are processed sequentially, and derive fundamental limits on the rate at which classical information can be transmitted using quantum states that decohere in time. Specifically, we model the sequential processing of qubits using a single server queue, and derive explicit expressions for the capacity of such a ‘queue-channel.’ We also demonstrate a sweet-spot phenomenon with respect to the arrival rate to the queue, i.e., we show that there exists a value of the arrival rate of the qubits at which the rate of information transmission (in bits/sec) through the queue-channel is maximised. Next, we consider a setting where the average rate of processing qubits is fixed, and show that the capacity of the queue-channel is maximised when the processing time is deterministic. We also discuss design implications of these results on quantum information processing systems.

I. INTRODUCTION

Quantum bits (or qubits) have a tendency to undergo rapid decoherence in time, due to certain fundamental physical phenomena. The manner and mechanism of such decoherence depends on the underlying physical implementation of the quantum state, the environment in which the quantum state evolves, and other physical factors such as temperature. Once a state decoheres, the information stored is lost either partially or completely, depending again on the underlying realizations and physical processes.

In this paper, we are concerned with *sequential processing of a stream of qubits* — for example, this can include transmitting, storing or performing gate operations on the quantum states. In this setting, we derive fundamental bounds on the rate at which information can be conveyed using quantum states that decohere in time.

When quantum states are prepared and then processed sequentially, it is reasonable to posit that there will inevitably be a non-zero ‘processing time,’ corresponding to each qubit, which in turn corresponds to a finite rate at which the qubits can be processed by the system. For example, when the qubit is prepared and transmitted as a photon polarization state, the rate at which a receiver can detect (and hence process) the photons is constrained by the average dead-time of the detectors, which is typically of the order of a tens of nanoseconds [1]. To consider another concrete example, superconducting Josephson junction based qubits have gate processing times ranging from a few tens of nanoseconds to a few hundreds of nanoseconds,

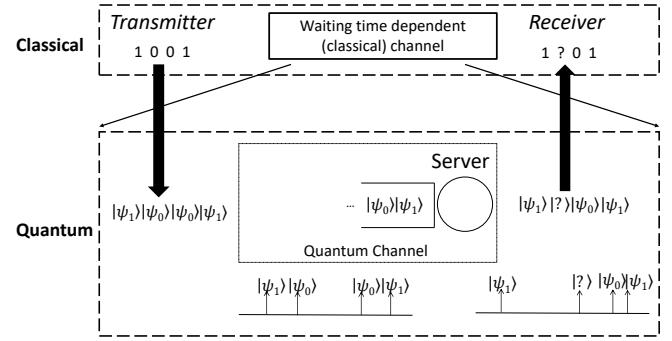


Fig. 1. Schematic of the queue-channel depicting the case of quantum erasure.

while their average coherence times are typically of the order of a few tens of microseconds [2, Table 2]. In such a scenario, the coherence time of each qubit is only about two to three orders of magnitude longer than the time it takes to process each qubit. This brings us to an interesting phenomenon, which does not seem to arise naturally in transmitting classical bits: as the qubits wait to be processed, they inevitably undergo decoherence, leading to errors.

More generally, we may consider a setting where qubits are prepared (or “arrive”) according to some random process at a particular rate, and are to be processed sequentially. Due to the non-zero processing time for each qubit, the arriving qubits will have to wait in sequence to be processed. The present paper focusses on obtaining a quantitative characterization of the above phenomenon. Specifically, we model the sequential processing of qubits using a single-server queue with average service rate μ . Now, suppose that the qubits ‘arrive’ at the queue according to a stationary random process of rate λ . Since the queue is stable if and only if $\lambda < \mu$, it is immediately clear that this system cannot process qubits at a rate higher than μ . A key question we address in this paper is as follows: Assuming for simplicity that each qubit is used to encode one classical bit, is it possible to transmit information through the above queue at a rate that is arbitrarily close to μ bits/sec?

In the case of classical bits, the answer is clearly in the affirmative. However, in the quantum case, we show that the answer turns out to be in the negative in general. Intuitively, when the arrival rate λ is very close to μ , the waiting time

for each qubit becomes very large. As a result, most of the qubits are likely to suffer decoherence, which leads to a higher probability of error.

Indeed, under physically well-motivated models for the decoherence of qubits with time, we derive explicit expressions for the capacity of the above ‘queue-channel’¹. In particular, we demonstrate a ‘sweet-spot’ phenomenon with respect to the arrival rate, i.e., we show that there exists a particular value of arrival rate $\lambda^* \in (0, \mu)$ at which the rate of information transmission (in bits/sec) through the queue-channel is maximised.

Next, for a given average rate μ of processing qubits, and Poisson arrivals of qubits, we prove that the channel capacity is maximised when the processing time of each qubit is deterministic. In other words, given a processing rate μ , the rate of information transmission is maximised by ensuring that the processing time is deterministic for each qubit.

Finally, we remark that similar waiting time dependent errors can also be observed in other emerging as well as classical systems. For example, due to the short-lived nature of human attention, the performance of a human deteriorates with the waiting time [4]. In this context, a waiting time dependent channel arises due to human impatience instead of quantum decoherence. This is particularly relevant to crowdsourcing. In the context of age of information [5], packets become useless (erased) after waiting in a queue for a certain duration — a scenario which also falls within the scope of the model we consider.

A. Related Literature

Gallager and Telatar initiated the area of multiple access queues in [6] which is the first published work at the intersection of queuing and information theory. Around the same time Anantharam and Verdú considered timing channels where information is encoded in the times between consecutive information packets, and these packets are subsequently processed according to some queueing discipline [7]. Due to randomness in the sojourn times of packets through servers, the encoded timing information is distorted, which the receiver must decode. In contrast to [7], we are not concerned with information encoded in the timing between packets — in our work, all the information is in the symbols.

An information theoretic notion of reliability of a queuing system with state-dependent errors was introduced and studied in [3], where the authors considered queue-length dependent errors motivated mainly by human computation and crowdsourcing.

B. Contributions

We focus on a simple queue-channel with waiting time induced *erasures* — specifically, we model the decoherence of qubits using a *quantum erasure channel* [8]. In the simplest M/M/1 setting, we explicitly characterize the capacity of the erasure queue-channel, and show that there is an optimal arrival rate $\lambda_{M/M/1} \in (0, \mu)$ at which the capacity of the

¹A terminology we borrow from [3].

queue-channel is maximised. Next, we generalize the above result to an M/GI/1 setting, and show a similar behaviour.

This result highlights an unusual interplay that exists between transmission rate and delay in the quantum case. Unlike in the classical case where we can obtain any rate that is arbitrarily close to the server rate (at the expense of delay), in the quantum case, it is desirable to operate away from the server capacity from the point of view of maximising capacity. This is because when qubits are sent faster than the optimal rate, the effective rate of information transmission actually *decreases*, due to a drastic increase in waiting time induced erasures.

While the above results characterize the optimal arrival rate of qubits for a fixed service distribution, one can also ask after the best service time distribution for *fixed* values of arrival and service rates. Indeed, we show that the capacity of the queue-channel is maximised when the service time distribution is deterministic. In other words, the M/D/1 queue maximises the queue-channel capacity, among all M/GI/1 queues. In certain physical realizations, there could be fundamental physical constraints that translate to an optimal gate processing rate of the qubits (see for example [9]). Our result offers an important design insight in such a scenario — the capacity is maximised when the gate processing time is deterministic across qubits, i.e., it is desirable to mitigate ‘jitter’ in the processing times.

II. SYSTEM MODEL

The model we study is depicted in Fig. 1. Specifically, a source generates a classical bit stream, which is encoded into qubits. These qubits are sent sequentially to a single server queue according to a stationary point process of rate λ . The server works like a FIFO queue with independent and identically distributed (i.i.d.) service times for each qubit. After getting processed by the server, each qubit is measured and interpreted as a classical bit. We refer to this system as a queue-channel, and characterize the classical capacity of this system (in bits/sec).

In order to capture the effect of decoherence due to the underlying quantum channel, we model the error probability as an explicit function of the waiting time W in the queue. For instance, in several physical scenarios, the decoherence time of a single qubit maybe modelled as an exponential random variable. In other words, the probability of a qubit error/erasure after waiting for a time W is given by $p(W) = 1 - e^{-\kappa W}$, where $1/\kappa$ is a characteristic time constant of the physical system under consideration [10, Section 8.3].

A. Queueing Discipline

We consider a continuous-time system. The service requirements are i.i.d. across qubits. The service time of the j th qubit is denoted S_j and has a cumulative distribution F_S . The average service rate of each qubit is μ , i.e., $E_{F_S}[S] = 1/\mu$. In the interest of simplicity and tractability, we assume Poisson arrivals, i.e., the time between two consecutive arrivals is i.i.d. with an exponential distribution with parameter λ . For stability

of the queue, we assume $\lambda < \mu$. For ease of notation let us assume $\mu = 1$. (Our results easily extend to general μ).

Let A_j and D_j be the arrival and the departure epochs of j th qubit, respectively and $W_j = D_j - A_j - S_j$ be the time that j th qubit waits in queue until its service begins.

B. Error Model

As the qubits wait to be served, they undergo decoherence, leading to errors at the receiver. This decoherence is modelled in general as a completely positive trace preserving map [10]. However, in this paper, we restrict ourselves to a rudimentary setting where we use a fixed set of orthogonal quantum states (say $|\psi_0\rangle$ and $|\psi_1\rangle$, corresponding to classical bits 0 and 1, as depicted in Fig. 1) to encode the classical symbols at the sender's side, and measure the qubits in some *fixed* basis at the receiver's end.

In general, the j th symbol $X_j \in \mathcal{X}$, is encoded as one of a set of orthogonal states $\{|\psi_{X_j}\rangle\}$ belonging to a Hilbert space \mathcal{H} of dimension $|\mathcal{X}|$. The noisy output state $|\tilde{\psi}_j\rangle$ is measured by the receiver in some fixed basis, and decoded as the output symbol $Y_j \in \mathcal{Y}$. This measurement induces a conditional probability distribution $\mathbf{P}(Y_j|X_j, W_j)$, which we can think of as an *induced classical channel* from \mathcal{X} to \mathcal{Y} .

An n -length transmission over the waiting time dependent queue-channel is denoted as follows. Inputs are $\{X_j : 1 \leq j \leq n\}$, channel distribution $\prod_j \mathbf{P}(Y_j|X_j, W_j)$, and outputs are $\{Y_j : 1 \leq j \leq n\}$.

Throughout, $Z^k = (Z_1, Z_2, \dots, Z_k)$ denotes a k -dimensional vector and $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n, \dots)$ denotes an infinite sequence of random variables. Information is measured in bits and log means logarithm to the base 2.

III. QUEUE-CHANNEL CAPACITY

We are interested in defining and computing the information capacity of the queue-channel, which is simply the capacity of the induced classical channel defined above. As mentioned earlier, we restrict ourselves to using a fixed set of orthogonal states to encode the classical symbols at the sender's side, and measuring in some fixed basis at the receiver's end. For this reason, the capacity of the induced classical channel is not the same as the *classical capacity of the quantum channel* resulting from the underlying decoherence model. The latter consideration is left for future work; see Section V.

A. Definitions

Let M be the message transmitted from a set \mathcal{M} and $\hat{M} \in \mathcal{M}$ be its estimate at the receiver.

Definition 1: An (n, \tilde{R}, T) code consists of the encoding function $X^n = f(M)$ and the decoding function $\hat{M} = g(X^n, A^n, D^n)$, where the cardinality of the message set $|\mathcal{M}| = 2^{n\tilde{R}}$, and for each codeword, the expected total time for all the symbols to reach to the receiver is less than T .

Definition 2: If the decoder chooses \hat{M} with average probability of error less than ϵ , that code is said to be ϵ -achievable. For any $0 < \epsilon < 1$, if there exists an ϵ -achievable code (n, \tilde{R}, T) , the rate $R = \frac{\tilde{R}}{T}$ is said to be achievable.

Definition 3: The information capacity of the queue-channel is the supremum of all achievable rates for a given arrival process with distribution F_A and is denoted by $C(F_A)$ bits per unit time.

We assume that the transmitter knows the arrival process statistics, but not the realizations before it does the encoding. However, depending on the application, the receiver may or may not know the realization of the arrival and the departure time of each symbol.

Proposition 1: The capacity of the queue-channel (in bits/sec) described in Sec. II is given by

$$C(F_A) = \lambda \sup_{\mathbf{P}(\mathbf{X})} \underline{I}(\mathbf{X}; \mathbf{Y} | \mathbf{W}), \quad (1)$$

when the receiver knows the arrival and the departure time of each symbol. On the other hand, when the receiver does not have that information, the capacity is,

$$C(F_A) = \lambda \sup_{\mathbf{P}(\mathbf{X})} \underline{I}(\mathbf{X}; \mathbf{Y}). \quad (2)$$

Here, \underline{I} is the usual notation for inf-information rate [11].

This result is essentially a consequence of the general channel capacity expression in [11]. Please see [12] for details. The following lemma which is used in proving Prop. 1 would be needed later.

Lemma 1: Under the assumptions in Sec. II, $\{W_j\}$ is a Markov process and has a unique limiting distribution π .

Proof: Follows from the stability results for GI/GI/1. ■

B. Remarks

Before proceeding further, we note the difference between the maximum symbol throughput (number of symbols processed per unit time) and the maximum information throughput (our notion of capacity) of the queuing system studied here. The symbol throughput is the maximum rate of arrivals for which the queue is stable and hence, increases with λ on $[0, \mu]$. On the other hand, the expression for ‘information throughput’ has λ as a multiplicative factor. However, this does not mean it increases with λ . In typical queuing systems, the average waiting time is increasing in λ . For quantum channels and other systems like crowd-sourcing and multimedia communication, service errors are more likely when waiting times are larger. Thus, increasing λ also negatively impacts the inf-information term in the capacity expression. Hence there is typically an information throughput-optimal $\lambda \in (0, \mu)$. This will be clear when we discuss some particular scenarios of interest.

The capacity expression in Prop. 1 does not provide clear insights into the behaviour of the system under different arrival and service statistics. Our main contribution in this paper lies in deriving a single letter capacity expression for queue-channels with waiting time induced *erasures*. A single letter capacity expression facilitates a clearer understanding of the effects of the arrival and service processes on the capacity. Analogous capacity results can also be derived for a class of M -ary symmetric queue-channels [12], but we omit discussing them here due to page constraints.

IV. ERASURE QUEUE-CHANNELS

Erasures channels are ubiquitous in classical as well as quantum information theory. We consider a quantum erasure channel [8] which acts on the j th state $|\psi_{X_j}\rangle$ as follows: $|\psi_{X_j}\rangle$ remains unaffected with probability $1-p(W_j)$, and is erased to a state $|\text{?}\rangle$ with probability $p(W_j)$, where $p : [0, \infty) \rightarrow [0, 1]$ is typically increasing. Such a model also captures the communication scenarios where information packets become useless (erased) after a deadline. For such an erasure channel, a single letter expression for capacity can be obtained.

Theorem 1: For the erasure queue-channel defined above, the capacity is $\lambda \log |\mathcal{X}| \mathbf{E}_\pi [1 - p(W)]$ bits/sec, irrespective of the receiver's knowledge of the arrival and the departure times of symbols.

Proof: The proof uses an upper-bound on $\underline{I}(\mathbf{X}; \mathbf{Y})$ in terms of unconditional sup-entropy rate and conditional inf-entropy rate and shows that the capacity expression is an upper-bound. On the other hand, using a similar lower-bound on $\underline{I}(\mathbf{X}; \mathbf{Y})$ we show that for a choice of distribution of $\{X_n\}$ (namely, i.i.d. uniform), $\underline{I}(\mathbf{X}; \mathbf{Y})$ is no smaller than the capacity expression. The fact that in the case of erasure channels the received symbol is either correct or erased (but never wrong) makes the knowledge of the arrival and departure times irrelevant. Finally, ergodicity of the queue is used in reducing n -symbol bounds for \underline{I} to single-letter bounds.

More precisely, using the properties of limit superior and limit inferior, we have

$$\underline{I}(\mathbf{X}; \mathbf{Y}|\mathbf{W}) \leq \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{W}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}, \mathbf{W}),$$

where $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{W})$ and $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}, \mathbf{W})$ are the lim-sup in probability of the sequences $\frac{1}{n} \log \frac{1}{\mathbf{P}(Y^n|W^n)}$ and $\frac{1}{n} \log \frac{1}{\mathbf{P}(Y^n|X^n, W^n)}$, respectively. By the channel model, given W_i and X_i , Y_i is independent of any other variable and hence, a product form would emerge. Also, note that $\mathbf{P}(Y_i|X_i, W_i) = p(W_i)$ if Y_i is an erasure, else, it is $1-p(W_i)$. Combining these observations we obtain

$$\begin{aligned} \frac{1}{n} \log \frac{1}{\mathbf{P}(Y^n|X^n, W^n)} &= -\frac{1}{n} \sum_{i=1}^n [\mathbf{1}(Y_i = \mathcal{E}) \log(1 - p(W_i)) \\ &\quad + \mathbf{1}(Y_i \neq \mathcal{E}) \log(p(W_i))], \end{aligned} \quad (3)$$

where \mathcal{E} represents erasure. By Lemma 1 the limit of the expression in (3) exists almost surely as a finite constant. Hence, this limit is the value of $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}, \mathbf{W})$. Note that for an erasure queue-channel this limit does not depend on the distribution of X^n .

Let us now consider upper-bounding $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{W})$. Let $N^\mathcal{E}$ be the set of indices for which $Y_i = \mathcal{E}$. Following standard conditional probability arguments, we get

$$\begin{aligned} &\log \frac{1}{\mathbf{P}(Y^n|W^n)} \\ &= - \sum_{i=1}^n [\mathbf{1}(Y_i = \mathcal{E}) \log(1 - p(W_i)) + \mathbf{1}(Y_i \neq \mathcal{E}) \log(p(W_i))] \\ &\quad - \log \mathbf{P}(\{Y_i : i \notin N^\mathcal{E}\} | \{Y_i \neq \mathcal{E} : i \notin N^\mathcal{E}\}, \{W_i : i \notin N^\mathcal{E}\}). \end{aligned} \quad (4)$$

Please see [12] for the detailed steps. As the almost sure limit of $\frac{1}{n} \sum_{i=1}^n [\mathbf{1}(Y_i = \mathcal{E}) \log(1 - p(W_i)) + \mathbf{1}(Y_i \neq \mathcal{E}) \log(p(W_i))]$ is equal to $\overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}, \mathbf{W})$, we only need to focus on the second term in (4).

Note that for an erasure channel, if Y_i is not an erasure, Y_i has the same value as that of X_i . So, for any joint distribution $\mathbf{P}_{\mathbf{X}}$ of input symbols:

$$\begin{aligned} &\frac{1}{n} (-\log \mathbf{P}(\{Y_i : i \notin N^\mathcal{E}\} | \{Y_i \neq \mathcal{E} : i \notin N^\mathcal{E}\}, \{W_i : i \notin N^\mathcal{E}\})) \\ &= -\frac{1}{n} \log \mathbf{P}_{\mathbf{X}}(\{Y_i : i \notin N^\mathcal{E}\} | \{W_i : i \notin N^\mathcal{E}\}) \\ &= -\frac{n - |N^\mathcal{E}|}{n} \frac{1}{n - |N^\mathcal{E}|} \log \mathbf{P}_{\mathbf{X}}(\{Y_i : i \notin N^\mathcal{E}\}) \end{aligned} \quad (5)$$

Note that in the limit, by Lemma 1, $\frac{|N^\mathcal{E}|}{n}$ converges almost surely to $\mathbf{E}[p(W)] < 1$. So, almost surely $n - |N^\mathcal{E}| \rightarrow \infty$. So as $n \rightarrow \infty$, $\frac{1}{n - |N^\mathcal{E}|} \log \mathbf{P}_{\mathbf{X}}(\{Y_i : i \notin N^\mathcal{E}\})$ can be seen as $\frac{1}{N} \log \mathbf{P}(X^N)$ for some large N . lim-sup in probability of this quantity is upper-bounded by $\log |\mathcal{X}|$. Thus we get an upper-bound of $(1 - \mathbf{E}[p(W)]) \log |\mathcal{X}|$ on \underline{I} .

To obtain a lower bound, we have

$$\underline{I}(\mathbf{X}; \mathbf{Y}|\mathbf{W}) \geq \underline{\mathbf{H}}(\mathbf{Y}|\mathbf{W}) - \overline{\mathbf{H}}(\mathbf{Y}|\mathbf{X}, \mathbf{W}), \quad (6)$$

where $\underline{\mathbf{H}}(\mathbf{Y}|\mathbf{W})$ is the lim-inf in probability of $\frac{1}{n} \log \frac{1}{\mathbf{P}(Y^n|W^n)}$. We have already derived an expression for the second term above. Then, using (4) and similar arguments which lead to (5) we obtain that the right hand side of (6) equals (5).

If we choose uniform and i.i.d. $\{X_i\}$, (5) almost surely converges to $(1 - \mathbf{E}[p(W)]) \log |\mathcal{X}|$. Thus, we derived a $\mathbf{P}_{\mathbf{X}}$ independent upper-bound on $\underline{I}(\mathbf{X}; \mathbf{Y}|\mathbf{W})$ which is matched by a particular choice of $\mathbf{P}_{\mathbf{X}}$. Thus, for the erasure channel

$$\sup_{\mathbf{P}_{\mathbf{X}}} \underline{I}(\mathbf{X}; \mathbf{Y}|\mathbf{W}) = (1 - \mathbf{E}[p(W)]) \log |\mathcal{X}|.$$

By multiplying with λ we obtain the capacity of this channel. See [12] for a complete proof. ■

This single letter capacity expression allows us to mine deeper insights on system design. It is well known in queuing that waiting time increases with increasing arrival rate. As $p(\cdot)$ is increasing, so is $\mathbf{E}_\pi[p(W)]$ in λ . Therefore, it is apparent from the single letter expression (in Theorem 1) that capacity may not be monotonic in λ . This raises an interesting question: is there an optimal λ at which the capacity is maximised? The answer to this question depends on the queuing dynamics. Therefore, we first attempt to understand it for the most fundamental queuing system in communication networks, the M/M/1 queue. Interestingly, for the M/M/1 queue, there exists a simple characterization of the capacity and the corresponding optimal arrival rate.

Theorem 2: The arrival rate that maximises the information capacity of the M/M/1 queue-channel is given by

$$1 - \arg \min_{u \in (0, 1)} u \left(1 + \tilde{p} \left(\frac{u}{1-u} \right) \right) \quad (7)$$

where for any $u > 0$, $\tilde{p}(u) := \int \exp(-ux)p(x)dx$ is the Laplace transform of $p(\cdot)$.

Proof: This proof uses the exponential waiting time distribution of M/M/1 queue to relate the capacity to Laplace transform of $p(\cdot)$.

It is known that the waiting time in M/M/1 is distributed as $\exp\left(\frac{1-\lambda}{\lambda}\right)$ for $\mu = 1$. Thus,

$$\begin{aligned}\mathbf{E}[p(W)] &= \int_0^\infty p(w) \frac{1-\lambda}{\lambda} \exp\left(\frac{1-\lambda}{\lambda}w\right) dw \\ &= \frac{1-\lambda}{\lambda} \tilde{p}\left(\frac{1-\lambda}{\lambda}\right).\end{aligned}$$

Thus, the capacity is given by $\lambda \left(1 - \frac{1-\lambda}{\lambda} \tilde{p}\left(\frac{1-\lambda}{\lambda}\right)\right)$.

So, the capacity maximising arrival rate is the one that maximises this expression:

$$\begin{aligned}&\arg \max_{\lambda \in (0,1)} \lambda \left(1 - \frac{1-\lambda}{\lambda} \tilde{p}\left(\frac{1-\lambda}{\lambda}\right)\right) \\ &\iff \arg \max_{\lambda \in (0,1)} \left(\lambda - (1-\lambda)\tilde{p}\left(\frac{1-\lambda}{\lambda}\right)\right) \\ &\iff 1 - \arg \max_{u \in (0,1)} \left(1 - u - u\tilde{p}\left(\frac{u}{1-u}\right)\right) \\ &\iff 1 - \arg \min_{u \in (0,1)} u \left(1 + \tilde{p}\left(\frac{u}{1-u}\right)\right).\end{aligned}$$

■

In the case of quantum erasure channels [8], decoherence of qubits over time gives rise to an interesting form for $p(\cdot)$, namely, $p(W) = 1 - \exp(-\kappa W)$, where κ is a physical parameter. A detailed quantum physical discussion on this can be found in [13]. A relation of this kind between waiting time and erasure is also relevant in multimedia communication with deadlines and in the context of age of information. In these scenarios, when deadlines or maximum tolerable age of information packets are unknown, the exponential distribution (being the most entropic) serves as a reasonably good stochastic model. Such a model is captured by the above form of $p(\cdot)$. Hence, for this particular form of $p(\cdot)$, it is important to understand the capacity behaviour explicitly.

Corollary 1: For an M/M/1 erasure queue-channel with $p(W) = 1 - \exp(-\kappa W)$, $F_A(x) = 1 - \exp(-\lambda x)$ and $F_S(x) = 1 - \exp(-x)$,

- (i) the capacity is given by $\frac{\lambda(1-\lambda)}{1-\alpha\lambda}$ bits/sec, and
- (ii) the capacity is maximised at

$$\lambda_{M/M/1} = \frac{1}{\alpha} \left(1 - \sqrt{1-\alpha}\right) = \frac{1}{1+\sqrt{1-\alpha}}$$

where $\alpha = \frac{1}{1+\kappa}$.

This result offers interesting insights into the relation between the information capacity and the characteristic time-constant of the quantum medium. In the case of decohering channels, a larger value of the decoherence exponent κ corresponds to a faster decoherence. We note that α decreases as κ increases and hence, $\lambda_{M/M/1}$ decreases as κ increases. This implies that when the qubits decohere more rapidly, the arrival

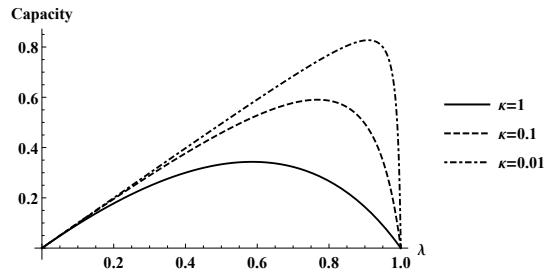


Fig. 2. The capacity of the M/M/1 queue-channel (in bits/sec) plotted as a function of the arrival rate λ for different values of the decoherence parameter κ .

rate that maximises the capacity is lower. In other words, when the coherence time is small, it is better to send at a slower rate to avoid excessive waiting time induced errors.

Fig. 2 depicts a capacity plot of the M/M/1 queue-channel (in bits/sec), as a function of the arrival rate λ for different values of the decoherence parameter κ . Since the service rate μ is taken to be unity, we note that a value of $\kappa = 0.01$ corresponds to an average coherence time which is two orders of magnitude longer than the service time — a setting reminiscent of superconducting qubits [2]. We also notice from the shape of the capacity curve for $\kappa = 0.01$ that there is a drastic drop in the capacity, if the system is operated beyond the optimal arrival rate $\lambda_{M/M/1}$. This is due to the drastic increase in delay induced decoherence as the arrival rate of qubits approaches the server capacity.

Next, we discuss the generalization of Corollary 1 to M/GI/1 queues. Specifically, a result similar to Corollary 1 also holds for M/GI/1 system for a different α , though unlike for the M/M/1 queue, the waiting time is not exponentially distributed.

Theorem 3: For an M/GI/1 erasure queue-channel with $p(W) = 1 - \exp(-\kappa W)$, $F_A(x) = 1 - \exp(-\lambda x)$, and a general F_S with $F_S(0) = 0$

- (i) the capacity is given by $\frac{\lambda(1-\lambda)}{1-\alpha\lambda}$ bits/sec, and
- (ii) the capacity is maximised at

$$\lambda_{M/GI/1} = \frac{1}{\alpha} \left(1 - \sqrt{1-\alpha}\right) = \frac{1}{1+\sqrt{1-\alpha}},$$

where $\alpha = \frac{1-\tilde{F}_S(\kappa)}{\kappa}$, and $\tilde{F}_S(u) = \int \exp(-ux)dF_S(x)$ is the Laplace transform of the service time distribution.

Proof: First, note that for the particular form of $p(\cdot)$ considered here, the capacity expression in Theorem 1 simplifies to $\lambda \mathbf{E}[\exp(-\kappa W)]$.

Next, using the Pollaczek-Khinchin formula for the M/GI/1 queue (with $\mu = 1$), we write

$$\mathbf{E}[\exp(-\kappa W)] = \frac{(1-\lambda)\kappa}{\kappa - \lambda(1-\tilde{F}_S(\kappa))} = \frac{1-\lambda}{1-\alpha\lambda},$$

where $\alpha = \frac{(1-\tilde{F}_S(\kappa))}{\kappa}$. Thus, the capacity is $\frac{\lambda(1-\lambda)}{1-\alpha\lambda}$ bits/sec, and the capacity maximising arrival rate is $\arg \max_{\lambda \in [0,1]} \frac{\lambda(1-\lambda)}{1-\alpha\lambda}$.

The objective function in the above optimization problem is concave in λ . This implies that the value of λ that maximises the capacity is the one at which the derivative of the capacity with respect to λ is zero. Taking the derivative, we obtain a quadratic function in λ which when equated to zero yields two solutions for λ : $\frac{1}{\alpha} \pm \frac{\sqrt{1-\alpha}}{\alpha}$. The only valid solution for which $\lambda \in [0, 1]$ is given by $\frac{1}{\alpha} - \frac{\sqrt{1-\alpha}}{\alpha}$. ■

The above results characterize an optimal λ for given arrival and service distributions. One can also ask after the best service distribution for a given arrival process and a fixed server rate. This question is of interest in designing the server characteristics like gate operations [9] or photon detectors in the case of quantum systems, as well as in the case of packet communication with age of information constraints. The following theorem is useful in such scenarios.

Theorem 4: For an erasure queue-channel with $p(W) = 1 - \exp(-\kappa W)$ and $F_A(x) = 1 - \exp(-\lambda x)$ at any λ the capacity is maximised by $F_S(x) = 1(x \geq 1)$, i.e., a deterministic service time maximises capacity, among all service distributions with unit mean and $F_S(0) = 0$.

Proof: As derived in the proof of Theorem 3, the capacity is

$$\frac{\lambda(1-\lambda)\kappa}{\kappa - \lambda(1 - \tilde{F}_S(\kappa))} = \frac{\frac{(1-\lambda)\kappa}{\lambda}}{\frac{\kappa-\lambda}{\lambda} + \tilde{F}_S(\kappa)}.$$

Thus, for any given λ , among all service distribution with unit mean, the capacity is maximised by that service distribution for which $\tilde{F}_S(\kappa)$ is minimised. For any service random variable S , by Jensen's inequality, we have $\tilde{F}_S(\kappa) = \mathbf{E}[\exp(-\kappa S)] \geq \exp(-\kappa \mathbf{E}[S])$. Therefore, $\tilde{F}_S(\kappa)$ is minimised by $S = \mathbf{E}[S]$, i.e., a deterministic service time. ■

Although we have considered only the erasure queue-channel in this paper, analogous results can be obtained for a more general class of channels which include binary (and M -ary) symmetric queue-channels. Please see [12] for details.

V. CONCLUDING REMARKS AND FUTURE WORK

In this paper, we used simple queue-channel models to characterize the capacity of channels with waiting time dependent errors. Though our main motivation stems from quantum communications, where we characterize the rate at which classical information can be transmitted using orthogonal quantum states that decohere in time, the model also captures scenarios in crowdsourcing and multimedia streaming.

We believe there is ample scope for further work along several directions. Firstly, it is important to move away from the restriction of using only orthogonal states at the encoder and a fixed measurement at the receiver, and allow for arbitrary superposition states at the encoder, and arbitrary measurements at the receiver. This would allow us to invoke the true classical capacity of the underlying (non-stationary) quantum channel, in terms of a quantity [14] analogous to the classical information rate. It remains an interesting technical challenge to obtain a formula for the queue-channel capacity in this general scenario, and identify channels for which the classical

coding strategy would still be optimal. Furthermore, we can also consider other widely studied quantum channel models, such as the phase damping and amplitude damping channels.

We have only considered uncoded quantum bits in this paper. We can also quantitatively evaluate the impact of using quantum codes to protect qubits from errors. Employing a code would enhance robustness to errors, but would also increase the waiting time due to the increased number of qubits to be processed. It would be interesting to characterise this tradeoff, and identify the regimes where using coded qubits would be beneficial or otherwise.

More broadly, we believe our work highlights the importance of explicitly modelling delay induced errors in quantum communications. As quantum computing takes strides towards becoming an ubiquitous reality, we believe it is imperative to develop processor architectures and algorithms that are informed by more quantitative studies of the impact of delay induced errors on quantum information processing systems.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewers for their helpful comments. Thanks also to Uday Khankhoje and Shweta Agrawal for that serendipitous lunch conversation which led to the model considered here.

REFERENCES

- [1] R. H. Hadfield, "Single-photon detectors for optical quantum information applications," *Nature photonics*, vol. 3, no. 12, p. 696, 2009.
- [2] G. Wendl, "Quantum information processing with superconducting circuits: a review," *Reports on Progress in Physics*, vol. 80, no. 10, p. 106001, 2017.
- [3] A. Chatterjee, D. Seo, and L. R. Varshney, "Capacity of systems with queue-length dependent service quality," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3950 – 3963, Jun. 2017.
- [4] D. Kahneman, *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
- [5] M. Costa, M. Codreanu, and A. Ephremides, "Age of information with packet management," in *Proc. 2014 IEEE Int. Symp. Inf. Theory*, Jun. 2014.
- [6] I. E. Telatar and R. G. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.
- [7] V. Anantharam and S. Verdú, "Bits through queues," *IEEE Trans. Inf. Theory*, vol. 42, no. 1, pp. 4–18, Jan. 1996.
- [8] C. H. Bennett, D. P. DiVincenzo, and J. A. Smolin, "Capacities of quantum erasure channels," *Physical Review Letters*, vol. 78, no. 16, p. 3217, 1997.
- [9] C. Ballance, T. Harty, N. Linke, M. Sepiol, and D. Lucas, "High-fidelity quantum logic gates using trapped-ion hyperfine qubits," *Physical review letters*, vol. 117, no. 6, p. 060504, 2016.
- [10] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.
- [11] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, Jul. 1994.
- [12] A. Chatterjee, K. Jagannathan, and P. Mandayam, "Qubits through queues: The capacity of channels with waiting time dependent errors," arXiv:1804.00906, 2018.
- [13] M. Grassl, T. Beth, and T. Pellizzari, "Codes for the quantum erasure channel," *Physical Review A*, vol. 56, no. 1, p. 33, 1997.
- [14] M. Hayashi and H. Nagaoka, "General formulas for capacity of classical-quantum channels," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1753–1768, 2003.

Detection of Vowel-Like Speech Using Variance of Sample Magnitudes

Nagapuri Srinivas, Gayadhar Pradhan and Puli Kishore Kumar

Department of Electronics and Communication Engineering

National Institute of Technology Patna, India

Email: {ns, gdp, pulkishorek}@nitp.ac.in

Abstract—Vowel, semi vowel and diphthong sound units are collectively referred to as vowel-like speech (VLS). VLS are dominant voiced regions in a given speech signal. Consequently, within a short-analysis frame the variance of sample magnitudes (VSM) is significantly higher for VLS when compared with other speech regions. In this work, a signal processing approach is proposed to robustly extract the VSM within an analysis frame. The VSM at each time instant is then non-linearly mapped (NLM) using negative exponential function to suppress the fluctuations. The NLM-VSM values are nearly constant and significantly less in magnitude for VLS than other speech, silence and noise regions. The NLM-VSM is used as a front-end feature for detecting the VLS in a given speech signal. The experimental results presented in this paper show that, for clean as well as noisy speech signals, the proposed feature outperforms some of the earlier reported features for the task of detecting VLS and corresponding onset and offset points.

I. INTRODUCTION

In the speech signal, semivowels are more similar with vowels when compared to other audio units [1]. When two vowels are pronounced consecutively in a single syllable, those two adjacent vowels are technically called as diphthong and are treated as a single sound unit [2]. In this work, vowel, semivowel and diphthong sound units are described as vowel-like speech (VLS) [1], [3]. VLS are more speaker specific and less degraded by environmental noises [1], [3]. In the earlier reported works, VLS are used for development of robust speech-based applications [1], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]. Performance of those systems critically depend on the accurate detection of VLS. In last two decades, several methods have been proposed for detecting vowel onset point (VOP), vowels and VLS [14], [15], [16], [3], [1]. VOP corresponds to the starting point of a vowel. In most of the existing vowel detection methods, semivowels and diphthongs are detected as vowels due to the similarities in their temporal and spectral characteristics [3], [17], [18]. The most of the vowel detection algorithms existing in the literature can not differentiate vowel with semivowels and diphthongs sounds. So those algorithms can also be employ for detection of VLS.

In a speech signal, the VLS are near-periodic, high energy and longer duration regions [2], [1], [3]. Considering these aspects, several front-end speech parameterization approaches capturing the behaviour of excitation and vocal-tract system responses have been reported for the detection of vowels and VOPs. Those front-end speech parameterization approaches

include the difference in energy of each of the peaks and their corresponding valleys in the the short-term discrete Fourier transform (DFT) magnitude spectrum [19], largest peaks in the DFT magnitude spectrum [14] and mel-frequency cepstral coefficients (MFCC), spectral energy in different frequency bands [18]. The feature representing excitation strength such as Hilbert envelope of the linear prediction (LP) residual [20] and rate of change of excitation strength extracted from the zero frequency filtered (ZFF) speech signal [1], [3] have also been explored. The zero-crossing rate, energy and pitch information of the speech signal [21], wavelet scaling coefficients of the speech signal [22], modulation spectrum energies [14], spectral energy present in the glottal closure regions [15], uniformity of the epoch intervals [23] and cumulative sum of the DFT magnitude spectrum of the non-local estimated speech signal [16] have also been used as the discriminating features. Several features have also been combined to represent the complementary information present in the vowels [14], [3], [1], [18]. Further, different statistical modeling methods like Hierarchical neural network, multilayer feed-forward neural network, auto-associative neural network [24] and hidden Markov model (HMM) are also explored. [18].

The VLS onset points (VLS-OPs) and offset points (VLS-EPs) are equally important for an accurate detection of VLS. For most of the VLS, speech samples around the VLS-EPs exhibit smaller transition when compared to those around the VLS-OPs. As a result, front-end features used for the detection of VLS-OPs deviate significantly at the VLS-EPs. In the case of approaches where VLS-OPs and VLS-EPs are used for detecting VLS, miss-detections are mainly due to poor detection of VLS-EPs [1], [25]. This is also true for the frame-based approaches where each analysis frame is classified either as VLS or non-VLS [18]. Furthermore, performance of the front-end speech parameterization techniques and statistical modeling methods, in terms of true detection and spurious rate, reduces significantly when the test signal is corrupted by ambient noise. Therefore, a front-end feature which is equally discriminative at the VLS-OPs and VLS-EPs and less affected by ambient noises is highly desirable for accurate detection of VLS in a speech signal.

In this paper, we propose a novel and robust technique that, exploits the variance of sample magnitudes (VSM) within a short-analysis frame as a front-end feature for detecting VLS-OPs, VLS-EPs and VLS. The experimental results presented in

this work show that the proposed feature is significantly robust towards the ill-effects of noise. To substantiate this claim, we have compared the proposed technique with two of the earlier reported font-end speech parametrization methods [15], [1]. In all the studied cases, the proposed approach is observed to outperform the existing ones.

The rest of the paper is organized as follows: In Section II, the proposed method for detecting VLS is described. The existing front-end speech parametrization methods used for performance comparison are discussed in Section III. The experimental results are presented in Section IV. Finally, the paper is concluded in Section V.

II. PROPOSED VLS DETECTION METHOD

It is very well known that, speech is a non-stationary signal. As a result, sample magnitudes within a short-analysis frame vary based on the underlying sound units. The VSM within the analysis frame for VLS is significantly higher as compared to other non-VLS, silence and noise regions. The change in VSM is also more pronounced at the VLS-OPs and VLS-EPs. The VSM directly computed from the speech signal deviates depending upon the type of noise and signal to noise ratio (SNR) of the input speech signal. If VSM is properly extracted, this can be used as a robust feature for discriminating the VLS in a given speech signal. The primary objective of this work is to robustly extract the VSM information from the speech signal and non-linearly map those to reduce the fluctuations for the task of detecting VLS-OPs, VLS-EPs and VLS. In the following sub-section, the proposed method for detecting VLS in a speech signal is described.

A. Proposed method

In the proposed method, the speech signal is processed through the following sequence of steps for detecting VLS-OPs, VLS-EPs and VLS. I.

- 1) The absolute value of the given speech signal ($|x(n)|$) is processed through a single pole filter whose transfer function is given as:

$$H(z) = \frac{1}{1 - rz^{-1}}. \quad (1)$$

The stability of the filter is ensured by placing the pole slightly inside the unit circle. In this work, r is selected as 0.999. The output of the filter is given by:

$$y(n) = ry(n-1) + |x(n)|. \quad (2)$$

In Eq. 2, absolute value of the given speech signal is taken to ensure that the output of the filter $y(n)$ as a growing polynomial of time.

- 2) Then, the VSM of the speech signal at each time instant is computed by using a frame of sample points of the filter output $y(n)$ as follows:

$$v(n) = \frac{1}{2l} \sum_{k=-l}^l (y(n+k) - \mu(n))^2 \quad (3)$$

where, $v(n)$ is the VSM of the speech signal and $\mu(n)$ is the mean of the filter output in the neighborhood of the sample points n , which is computed as follows:

$$\mu(n) = \frac{1}{2l+1} \sum_{k=-l}^l y(n+k) \quad (4)$$

where, $2l+1$ is the length of the sample neighborhood.

- 3) The transitions in VSM at VLS-OPs and VLS-EPs differs depending on the context of speech signal. In noisy speech signal VSM is also fluctuates depending on the type of noise and SNR. To reduce the fluctuations and enhance the transitions at the VLS-OPs and VLS-EPs, VSM at each time instant is non-linearly mapped using a negative exponential as follows:

$$v_{nl}(n) = \exp\left(-\frac{v(n)}{c}\right) \quad (5)$$

where, $v_{nl}(n)$ represents the non-linearly mapped VSM (NLM-VSM) and c is a real constant.

- 4) The notable transition points in the NLM-VSM are found by convolving it with a 100 ms long first-order difference of Gaussian window (FODG) having a standard deviation one sixth of the window size. In the convolved output, termed as the *VLS detection evidence*, the valleys and peaks correlate with the VLS-OPs and VLS-EPs, respectively. The VLS-OPs and VLS-EPs in the *VLS detection evidence* are found by calculating the sum of the magnitude of the valleys and the corresponding peaks. If sum of the magnitudes is above a predefined threshold, the valleys and corresponding peaks are hypothesized as the VLS-OPs and VLS-EPs, respectively. The regions between them are selected as the VLS.

The proposed approach for detecting VLS-OPs, VLS-EPs and VLS is shown in Fig. 1. The filter output for a segment of clean speech (Fig. 1 (a)) as well as for the same speech segment degraded by 0 dB white noise (Fig. 1 (f)) are shown in Fig. 1 (b) and Fig. 1 (g), respectively. It is evident from the figures that, the change in filtered output signal $y(n)$ at each time instant depends on the magnitude of the sample value $x(n)$ which, in turn, depends on the underlying sound unit. The VSM for clean and 0 dB white noise degraded speech signal are shown in Fig. 1 (c) and Fig. 1 (h), respectively. As discussed earlier, the VSM is more within the VLS as compared to other sound units. It can also be observed that, there is significant transitions present in VSM values at the VLS-OPs and the VLS-EPs. The NLM-VSM for clean as well as 0 dB white noise degraded speech are shown in Fig. 1 (d) and Fig. 1 (i), respectively. The NLM-VSM values are significantly less in magnitude and maintains constant values within the VLS. Furthermore, the fluctuation present in VSM is suppressed and the transitions at VLS-OPs and VLS-EPs are enhanced in NLM-VSM. The *VLS detection evidence* for clean and noise degraded speech signal are given in Fig. 1 (e) and Fig. 1 (j), respectively. From the *VLS detection evidences*, it may be observed that, the detected VLS (solid red lines)

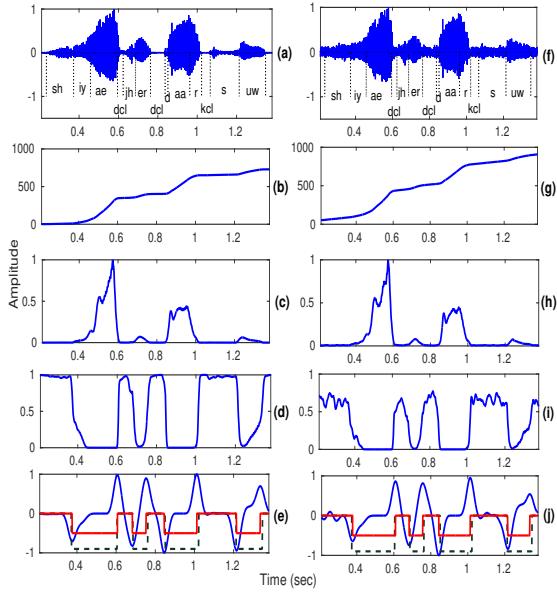


Fig. 1. Plots illustrating the proposed approach for detecting VLS-OPs, VLS-EPs and VLS in a given speech signal under clean and noisy conditions.(a) A segment of speech taken from the TIMIT database with reference marking of sound units as given in the database. (b) Single pole filter output for clean speech. (c) VSM for clean speech (d) NLM-VSM for clean speech (e) VLS evidence for detecting VLS-OPs, VLS-EPs and VLS with reference (black dash-dot lines) and detected (red solid line) VLS. (f) to (j) The same segment of the speech signal, filter output, VSM, NLM-VSM and VLS evidence, respectively, after adding 0 dB white noise.

nearly match to the reference ones (black dash-dot lines). By comparing the detected and reference VLS for the clean and noisy speech signal, it may be noted that the proposed feature is highly robust towards the ill-effects of environmental noise.

III. EXISTING METHODS FOR VLS DETECTION

For performance comparison two existing methods are explored [15], [1]. The approach presented in [1] was originally proposed for the detection of VLS-OPs, VLS-EPs and VLS where as the approaches presented in [15] were originally proposed for the detection of VOPs. In the first approach, the change in strength of excitation (CSE) derived from the ZFF speech signal was employed as the front-end feature for detecting VLS [1]. The positive zero crossings of the ZFF signal corresponds to the glottal closure instants (GCIs) [26]. The first-order difference of the ZFF speech signal at the GCIs corresponds to the strength of excitation [27]. As demonstrated in [1], the absolute value of the second order difference of ZFF speech signal represents the CSE. The CSE is significantly higher at the VLS-OPs and VLS-EPs as compared to the transitions of other sound units [1]. The fluctuations in the CSE was then smoothed by applying moving average filtering over a 50 ms window. The transition points at the VLS-OPs and the VLS-EPs in the smoothed CSE were detected by convolving it with 100 ms and 200 ms long FODGs, respectively. The VLS-OPs and VLS-EPs were detected independently. Then, one to one assignment of those points was done by using a

two step algorithm. The regions laying between a VLS-OP and corresponding VLS-EP was hypothesized as the VLS. In rest of the paper this approach is termed as *CSE-ZFF*.

In the second method [15], the GCIs were perceived from the ZFF speech signal. Anchoring the GCIs, short-term DFT magnitude spectrum was evaluated for the speech samples present in the 30% of glottal cycle. The spectral energies for those regions in the frequency band of 500-2500 Hz were used as the feature. The feature was further smoothed over 50 ms regions to suppress the variations. Next, the slope values were computed using the first order difference and the peaks having lower slope values were eliminated. The desired peaks are then locally enhanced by finding corresponding zero crossings. The significant transitions in the enhanced feature were detected by convolving it with 100 ms long FODG. In the convolved output, the peaks and valleys corresponds to the VLS-OPs and VLS-EPs, respectively. The VLS-OPs and VLS-EPs in the convolved output were detected by considering the sum of the magnitude of the peaks and the corresponding valleys. If sum of the magnitudes is above a predefined threshold, the peaks and corresponding valleys were hypothesized as the VLS-OPs and VLS-EPs, respectively. The regions between them were selected as the VLS. In rest of the paper this approach is termed as *SE-GCI*.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this section, the experimental setup for the various studies performed in this work are presented. This section also tabulates various experimental results.

A. Speech data

In this work, TIMIT database [28] is used for performance evaluation. Test dataset consists of 200 randomly selected utterances, equally divided between male and female speakers. Performance of the proposed and existing methods are evaluated on this datasets. A development set consisting of 50 utterances is used for optimizing the tunable parameters of the proposed and existing methods. To simulate noisy test conditions, three different noises namely White, Factory and Babble noise from the NOISEX-92 database [29] are added to the speech files. The energy level of the noise is varied so that the SNR of the noisy speech is either 0, 5, 10 or 15 dB.

B. Selection of tunable parameters for proposed approach

The estimation of VSM depends on two tunable parameters. Those are length of the analysis frame (l) and value of the constant (c) used as an normalization factor in the negative exponential function. The constant c , controls the degree of smoothing applied to the VSM. This, in turn, will reduce the variations present in the VLS and enhances the transitions present at the VLS-OPs and VLS-EPs. In this study the value of l and c are selected as 120 and 0.02, respectively. These values are selected empirically using the development dataset. For all the experiments, these values are kept constant to simulate a realistic testing conditions.

TABLE I
PERFORMANCES OF DIFFERENT APPROACHES EXPLORED/PROPOSED IN THIS PAPER FOR DETECTING VOWELS IN A GIVEN SPEECH SIGNAL UNDER CLEAN AND NOISY TEST CONDITIONS.

Speech Data Type	Method	VLSOP			VLSEP			VLS				
		IR in \pm ms			SR	SIR in \pm ms			SSR	SSR_s	SSR_{ns}	
		10	20	30		10	20	30				
Degraded by noise												
10 dB SNR	CSE-ZFF	63.27	73.20	80.75	11.66	61.76	72.01	79.17	14.25	74.21	13.08	13.20
	SE-GCI	61.17	73.09	80.15	13.51	58.43	66.26	73.77	15.51	66.89	10.85	14.52
	Proposed	73.18	79.61	82.81	9.39	72.46	78.48	81.87	11.55	84.89	10.84	6.54
5 dB SNR	CSE-ZFF	61.53	72.59	80.18	17.91	58.05	71.01	78.37	17.25	73.65	13.67	14.04
	SE-GCI	60.44	72.58	79.94	17.98	55.32	66.04	72.73	19.09	66.47	11.48	16.66
	Proposed	70.00	78.50	80.40	13.49	68.99	75.86	80.52	14.70	84.49	14.63	7.31
0 dB SNR	CSE-ZFF	58.54	71.64	79.33	18.78	55.23	69.22	77.34	18.72	72.94	14.41	16.67
	SE-GCI	59.34	71.82	79.56	20.28	51.15	65.51	73.45	21.29	66.50	11.09	17.40
	Proposed	68.95	76.99	79.23	14.67	67.55	74.82	79.11	16.18	83.54	17.11	9.08

C. Metrics for performance evaluation

The accuracy of the explored and proposed approaches for the task of detecting VLS-OPs and VLS-EPs are measured using the manual markings given in the database as the reference. The performance is measured using the following metrics:

- *Identification rate (IR)*: The percentage of the reference VLS-OPs/VLS-EPs that match with the detected VLS-OPs/VLS-EPs within the pre-defined deviation (in ms).
- *Spurious rate (SR)*: The percentage of detected VLS-OPs/VLS-EPs, which are detected outside the reference vowel regions.

The VLS detection performances of the explored and proposed approaches are measured using the following parameters:

- *Sample identification rate (SIR)*: The percentage of reference VLS samples that exactly match with the detected VLS samples.
- *Sample spurious rate (SSR)*: The percentage of detected VLS samples that lie outside the reference VLS. For a more detailed analysis, spurious detection is further broken into following two categories: i.
 - 1) SSR_s : The percentage of the detected VLS samples that match with non-VLS samples.
 - 2) SSR_{ns} : The percentage of the detected VLS samples that match with non-speech samples.

D. Detection of VLS-OPs and VLS-EPs

The performances of the proposed and existing methods for the task of detecting VLS-OPs and VLS-EPs in clean as well as noisy speech signal in terms *IR* and *SR* are given in Table I. From these results, it is evident that, the proposed feature provides better *IR* as well as *SR* when compared with the existing methods. The *IR* improvements observed in the

lower deviation (± 10 ms) are significantly large. Furthermore, the *IR* improvements in case of VLS-EPs is more as compared to the VLS-OPs. This shows that, the proposed feature is equally discriminative both at the VLS-OPs and VLS-EPs. The reduction in *IR* for the noisy test conditions is also less for the proposed approach.

E. Detection of VLS

The detection accuracy of VLS for the proposed as well as existing methods are summarized in Table I. For the clean speech, the *SIR* obtained by the proposed method is significantly higher than all the existing approaches explored in this study. The SSR_s and SSR_{ns} are also less in case of the proposed method. Furthermore, the reduction in *SIR* for the noisy test conditions is less for the proposed approach. This shows that the proposed approach is capable of detecting the VLS even in severely degraded conditions. These experiments highlights the efficiency and robustness of the proposed method in suppressing the ill-effects of the environmental variation for the task of detecting VLS to a large extent.

V. SUMMARY AND CONCLUSIONS

In this work, the VSM within an analysis frame is computed using the output of a single pole filtered speech. The VSM for the VLS frames are significantly higher than the non-vowel and silence regions. Further, the VSM at each time instant is nonlinearly mapped using negative exponential function to minimize the variations within the long-duration VLS. The non-linearly mapped VSM is used as a feature for detecting VLS in a speech signal. The proposed approach is also suitable for detecting VLS-OPs and VLS-EPs. For proper validation, the proposed method is compared with three existing approaches. The experimental results presented in this work show that the proposed feature provides significantly improved performance under clean as well as noisy test conditions when compared with the existing methods.

REFERENCES

- [1] G. Pradhan and S. M. Prasanna, "Speaker verification by vowel and non-vowel like segmentation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 4, pp. 854–867, Apr. 2013.
- [2] K. N. Stevens, *Acoustic Phonetics*. The MIT Press Cambridge, Massachusetts, London, England, 2000.
- [3] S. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded conditions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2552–2565, Nov. 2011.
- [4] N. Almaadeed, A. Aggoun, and A. Amira, "Text-independent speaker identification using vowel formants," *Journal of Signal Processing Systems*, vol. 82, no. 3, pp. 345–356, May 2015.
- [5] N. Fakotakis, A. Tsopanoglou, and G. Kokkinakis, "A text-independent speaker recognition system based on vowel spotting," *Speech Communication*, vol. 12, no. 1, pp. 57–68, Mar. 1993.
- [6] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Applied Soft Computing*, vol. 27, pp. 231–239, Feb. 2015.
- [7] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Spotting and recognition of consonant-vowel units from continuous speech using accurate detection of vowel onset points," *Circuits, Systems, and Signal Processing*, vol. 31, no. 4, pp. 1459–1474, Feb. 2012.
- [8] ———, "Improved consonant-vowel recognition for low bit-rate coded speech," *International Journal of Adaptive Control and Signal Processing*, vol. 26, no. 4, pp. 333–349, Oct. 2011.
- [9] S. P. Panda and A. K. Nayak, "Automatic speech segmentation in syllable centric speech recognition system," *Int. J. Speech Technol.*, vol. 19, no. 1, pp. 9–18, Nov. 2016.
- [10] C. Themistocleous, "Dialect classification using vowel acoustic parameters," *Speech Communication*, vol. 92, pp. 13–22, Sep. 2017.
- [11] S. Deb and S. Dandapat, "Emotion classification using segmentation of vowel-like and non-vowel-like regions," *IEEE Transactions on Affective Computing*, vol. 99, pp. 1–15, Jul. 2017.
- [12] K. S. Rao and B. Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Communication*, vol. 51, no. 12, pp. 1263–1269, Dec. 2009.
- [13] K. S. Rao and A. K. Vuppala, "Non-uniform time scale modification using instants of significant excitation and vowel onset points," *Speech Communication*, vol. 55, no. 6, pp. 745–756, Jul. 2013.
- [14] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 556–565, Mar. 2009.
- [15] A. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1894–1903, Apr. 2012.
- [16] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Non-local estimation of speech signal for vowel onset point detection in varied environments," in *Proc. INTERSPEECH*, Aug. 2017, pp. 429–433.
- [17] ———, "Exploring different acoustic modeling techniques for the detection of vowels in speech signal," in *Proc. National Conf. on Communication (NCC)*, Mar. 2016, pp. 1–5.
- [18] ———, "Improvements in the detection of vowel onset and offset points in a speech sequence," *Circuits, Systems, Signal Process.*, vol. 36, pp. 1–26, Sept. 2016.
- [19] D. J. Hermes, "Vowel onset detection," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 866–873, Feb. 1990.
- [20] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation source information," in *Proc. Interspeech*, Sept. 2005, pp. 1133–1136.
- [21] J. Wang, C. Hu, S. Hung, and J. Lee, "A hierarchical neural network based C/V segmentation algorithm for Mandarin speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2141–2146, Sep. 1991.
- [22] J. H. Wang and S. H. Chen, "A C/V segmentation algorithm for Mandarin speech using wavelet transforms," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, vol. 1, Mar. 1999, pp. 417–420.
- [23] A. K. Vuppala, K. S. Rao, and S. Chakrabarti, "Improved vowel onset point detection using epoch intervals," *AEU-Int. J. Electron. Commun.*, vol. 66, no. 8, pp. 697–700, Aug. 2012.
- [24] J. Rao, C. C. Sekhar, and B. Yegnanarayana, "Neural network based approach for detection of vowel onset points," in *Proc. Int. Conf. Adv. Pattern Recognition Digital Tech.*, vol. 1, Dec. 1999, pp. 316–320.
- [25] J. Yadav and K. S. Rao, "Detection of vowel offset point from speech signal," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 299–302, Apr. 2013.
- [26] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.
- [27] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, Jun. 2009.
- [28] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Linguistic Data Consortium, Dec. 1993, vol. 33.
- [29] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.

Detection of Vowels in Speech Signals Degraded by Speech-Like Noise

Avinash Kumar Dept. of ECE, NIT Patna Patna 800005, India k.avinash@nitp.ac.in	Sarmila Garnaik Dept. of EEE, VSSUT Odisha 768018, India sgarnaik_eee@vssut.ac.in	Ishwar Chandra Yadav Dept. of ECE, NIT Patna Patna 800005, India ishwarchy.ec15@@nitp.ac.in	Gayadhar Pradhan Dept. of ECE, NIT Patna Patna 800005, India gdp@nitp.ac.in
---	--	--	--

Syed Shahnawazuddin
Dept. of ECE, NIT Patna
Patna 800005, India
s.syed@nitp.ac.in

Abstract—Detecting vowels in a noisy speech signal is a very challenging task. The problem is further aggravated when the noise exhibits speech-like characteristics, e.g., babble noise. In this work, a novel front-end feature extraction technique exploiting variational mode decomposition (VMD) is proposed to improve the detection of vowels in speech data degraded by speech-like noise. Each short-time analysis frame of speech is first decomposed into a set of variational mode functions (VMFs) using VMD. The logarithmic energy present in each of the VMFs is then used as the front-end features for detecting vowels. A three-class classifier (vowel, non-vowel and silence) with acoustic modeling based on long short-term memory (LSTM) architecture is developed on the TIMIT database using the proposed features as well as mel-frequency cepstral coefficients (MFCC). Using the three-class classifier, frame-level time-alignments for a given speech utterance are obtained to detect the vowel regions. The proposed features result in significantly improved performance under noisy test conditions than the MFCC features. Further, the vowel regions detected using the proposed features are also quite different from those obtained through the MFCC. Exploiting the aforementioned differences, the evidences are combined to further improve the detection accuracy.

Index Terms: Vowel, speech-like noise, variational mode decomposition, variational mode function.

I. INTRODUCTION

The instants of time where a vowel sound starts within a speech segment are referred to as vowel onset points (VOPs) while vowel end points (VEPs) are those time instants where the vowel ends [1]–[4]. The vowels correspond to those region of speech signal that are periodic in nature and are of larger amplitude and longer duration [2], [5]. The excitation source and the vocal tract system response are predominantly reflected within the vowel regions. These aspects of speech production have been exploited in the existing methods for detecting the vowel regions and their corresponding VOPs and VEPs [2]–[4], [6]. The vowel regions, the VOPs and VEPs are employed in extracting features that are robust towards environmental degradations.

It is well known that, the transition characteristics of the vowels vary with the linguistic context of the spoken utter-

ance [7], [8]. For example, the transition from a fricative to vowel is completely different from that for a semivowel to vowel transition. Due to the similarities in the production characteristics of the vowels and semivowels, accurate detection of vowel are observed to be very challenging. The detection of vowel has many application in the field of speech recognition, speaker recognition, speech analysis, detecting consonant-vowel units, emotion classification, prosody modification, speech segmentation and keyword spotting [1]–[3], [9], [10]. Considering these applications, several front-end speech parameterization approaches have been proposed for the detection of vowels in speech signal [2], [3], [11]–[13]. The existing methods based on the transition characteristics are generally threshold dependent. In general, the vowel are detected by convolving the features characterizing the temporal variations with a first order Gaussian difference (FOGD) operator within a region that is 100 ms in duration [2]–[4], [6]. The selection of 100 ms region is an empirical choice since smaller region gives more spurious detection while larger regions decrease the accuracy for detection of vowel. This is so because, most of the weak transitions are detected as vowels and smoothed out correspondingly.

The task of detecting vowels in noisy condition is even more challenging. In general, the features associated with the detection of vowels deviate significantly depending on the kind of noise and input signal to noise ratio (SNR) [8], [11]. Therefore, the threshold dependent approaches fail to correctly classify vowels from non-vowels. To address this shortcoming, we have presented a vowel detection system in this work that is threshold independent. The developed system exploits statistical modeling of the vocal tract system. Since the effect of ambient noise on energy and transitions at the VOPs and VEPs varies in different frequency bands, data-adaptive frequency band energies are computed and those are used as the front-end features in this work.

In order to accurately detect the vowel regions, acoustic modeling based on long short-term memory (LSTM) [14] recurrent neural network is employed in this study. For detecting

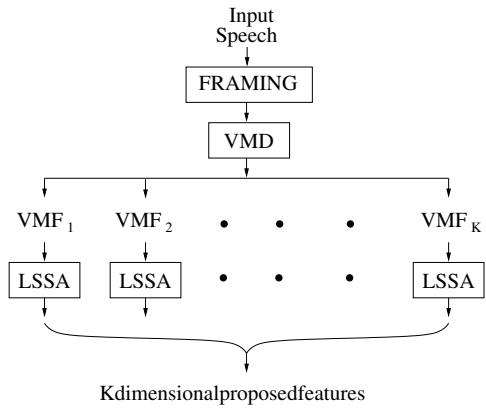


Fig. 1: Block diagram representing the process of extracting the proposed feature employing log sum squared amplitude (LSSA) operation.

vowels, a three-class classifier (vowel, non-vowel and silence) is developed. The speech sound units excluding the vowels are termed as non-vowels in this study. Two separate classifiers are developed using the proposed features and the conventional mel-frequency cepstral coefficients (MFCC) [15] on the TIMIT database [16]. To obtain the frame-level alignments, the given test speech data is forced-aligned against the corresponding acoustic models. The frame-level alignments are, in turn, used to detect the vowel regions. To determine the impact of semivowel and nasal sound units on the detection of vowel regions, the correctly detected and spurious vowel regions are also analyzed in detail. During our experimental evaluations it was observed that, the vowel regions detected using the MFCC and proposed features are quite different. Motivated by that, a novel scheme to combine the obtained evidences is also proposed in this work. Combining the evidences is found to improve the accuracy with which the vowel samples are detected.

The rest of the paper is organized as follows: The proposed approach for computing data-adaptive frequency band energies is described in Section II. In Section III, the experimental evaluations and a detail analysis of vowel detection is explained. Finally, the paper is concluded in Section IV.

II. PROPOSED FRONT-END FEATURES

The energy of different sound units in a speech sequence within different frequency bands differ significantly. For example, the energy in the case of most of the nasals is confined in between 0-500 Hz while that in the case of vowel sound units, the range of frequency is 500-2500 Hz [3], [4]. This fact is exploited to decompose an analysis frame of speech signal into different bands of frequency. Variational mode decomposition (VMD) [17] is used to adaptively decompose the speech signal in this study. VMD is a technique to non-recursively decompose a sequence into a discrete number of band-limited sub-sequences referred to as modes. Each of the modes, in turn, has a compact frequency support around a center frequency. In order to identify these modes, a constrained optimization routine exploiting alternating direction method of multipliers is employed. During optimization step, the sum of

the bandwidth of modes is minimized subject to the condition that the sum of the modes exactly reconstructs the original signal. In general, the number of modes is fixed beforehand.

Let $x^i(n)$ represent the i th short-time frame derived from a given speech utterance using a frame-size of 20 ms and a frame-shift of 10 ms. Further, let N^i represent the number of samples in a 20 ms frame and K be the number of variational mode functions (VMFs) into which the short-time frames are decomposed. Let, $x_k^i(n)$ denote the k th VMF for the i th short-time frame. The proposed features \mathcal{F}^i for the i th frame are then derived from the VMFs using the following equation:

$$\mathcal{F}^i(k) = \log \left[\sum_{n=0}^{N^i} \{x_k^i(n)\}^2 \right], \quad \text{where } k = 1, 2, \dots, K. \quad (1)$$

Thus, a K -dimensional feature vector is obtained for each of the short-time frames. In this study, the number of VMFs (K) is chosen to be twelve. The log sum squared amplitude (LSSA) operation performed to derive the proposed features is similar to the log-compression applied on the mel-filterbank warped energies in the case of MFCC features. The primary objective here is to capture information that is present in the energy of the VMFs. The energy in the VMFs will vary depending on the underlying sound units. The overall procedure for computing the proposed data-adaptive features is summarized in Figure 1.

III. EXPERIMENTAL EVALUATION

A. Experimental setup

System development as well as evaluation reported in this paper was performed using TIMIT acoustic-phonetic speech corpus [16]. The TIMIT database was first split into non-overlapping train, test and development sets following the standard Kaldi recipe. The train set comprised of 3696 utterances from 462 speakers. A set of 192 utterances from 24 speakers was used as the test set. The development set, on the other hand, consisted of 400 utterances from 50 speakers. The development set was used for optimizing the system parameters. The sentence level transcription available with the database was modified to represent all the possible vowels as a single class (V). Similarly, the transcriptions corresponding to non-vowel sound units were clustered together to define the second class (NV). The silence, short-pause as well as other filler units represented the third class (sil). Using the training data and the modified transcription, a three-class classifier (V, NV, sil) was developed using the Kaldi speech recognition toolkit [18]. For evaluating the performance under noisy conditions, babble noise collected from NOISEX-92 database [19] was added to the test set. All the experimental studies reported in this paper were performed on narrow-band speech data (i.e., sampled at 8 kHz rate) in order to simulate telephone-based speech interface.

To compute the base MFCC features, the speech data was first split into short-time frames using overlapping Hamming windows of duration 20 ms with 50% overlap. The 13-dimensional base MFCC features ($C_0 - C_{12}$) were then

TABLE I: Classification error rates (CERs) for the LSTM-based three-class classifiers developed using the MFCC as well as proposed features.

Data set	Data type	CERs (in %)	
		MFFC	Prop. feat.
Test	Clean	12.61	13.48
	BN 10dB	17.80	18.19
	BN 5dB	25.40	22.24
	BN 0dB	48.06	26.40

extracted using a 23-channel mel-filterbank for each of the short-time frames. Next, the base features were spliced in time with context size being 9 frames (± 4). Time-splicing resulted in 117-dimensional vectors. Linear discriminant analysis (LDA) was used to project the 117-dimensional vectors to 40-dimensional subspace. This was followed by further de-correlation using maximum likelihood linear transformation (MLLT). In addition to these, cepstral mean and variance normalization (CMVN) as well as speaker normalization using feature-space maximum likelihood linear regression (fMLLR) were also employed. As in the case of MFCC features, the 12-dimensional base excitation features were time-spliced first. Next, using LDA, MLLT, CMVN, and fMLLR, the 40-dimensional feature vectors were derived.

LSTM-based acoustic modeling was used for developing the said vowel-non-vowel-silence detection system. Before learning the LSTM system parameters, the final 40-dimensional fMLLR-normalized features were spliced over 4 frames to the left and right of the frame being analyzed. The LSTM-based acoustic models were trained with 4 hidden layers each having 1024 nodes. The dimension of the LSTM cell was chosen as 1024. The number of epochs used for LSTM training was set to 5 while the initial and final learning rates were selected to be 0.005 and 0.0005, respectively.

B. Evaluation results

The classification error rates (CERs) for the three-class classifier developed using MFCC features are given in Table I. The CERs were computed in the same way as word error rates are determined with the possible words being V, NV and sil. The CERs for the proposed features are also given in Table I. From the enlisted CERs, it can be observed that the classifier developed using the proposed features is inferior to that trained on MFFC features when the test data is devoid of noise. On the other hand, when speech-like noise i.e., babble noise is added to the test set, the proposed features result in lesser CERs especially for low SNR cases. In the following, we present a more detailed analysis on the detection of vowel regions withing a speech segment. Next, we present a simple technique to enhance the accuracy with which vowel regions are detected.

1) *Detection of vowel regions:* In order to detect the vowel regions, the given test utterance was forced-aligned with respect to the LSTM-based acoustic models. Forced-alignment was performed under the constraints of the first-pass hypothesis. This process results in generating frame-level alignments

which are required to detect the vowel regions. Using the detected vowel regions, VOPs and VOPs are determined by finding the starting and the ending points, respectively. In order to gauge the accuracy with which the vowel regions are detected, the manual markings provided with the database are used as the reference. The performances are then measured using the following two metrics:

- *Identification rate (IR):* The percentage of reference vowel samples that exactly match with the detected vowel samples.
- *Spurious rate (SR):* The percentage of detected vowels that lie outside the reference vowels.

To perform a more detailed analysis, spurious detection is further divided into:

- (i) *SR for semivowels:* The percentage of the detected vowel samples that exactly match with reference semivowel samples.
- (ii) *SR for others:* The percentage of the detected vowel samples that match with other speech samples (excluding vowels and semivowels).

The *IR* and *SR* for the LSTM-based three-class classifiers developed using the MFCC and proposed features are given in Table II. Both *IR* and *SR* obtained using the proposed features are observed to be inferior to those obtained through MFCC when the test data is clean. When speech-like noise is added to the test set, *IR* and *SR* obtained using the proposed features are much better than those obtained using MFCC. At the same time, due to inherent differences between the MFCC and proposed features, the vowel regions detected using the two types of features turn out to be very different as illustrated in Figure 2. When compared to reference markings (black dashed lines), the starting and ending points for the vowel regions detected using MFCC (pink dotted dashed lines) and proposed features (black dotted lines) happen to be very different. Consequently, the two evidences can be combined in an effective manner to accurately determine the boundaries for the vowel regions. Motivated by this fact, a novel technique for combining the evidences is discussed in the following.

2) *A novel technique for combining evidences:* In order to enhance the accuracy with which the vowel regions are detected, we employed a weighted combination of evidences obtained using MFCC and proposed features. From earlier presented evaluations, it is evident that the proposed features are better than the conventional MFCC features under noisy test conditions. Consequently, while combining the evidences, the evidence obtained using the proposed features is given a higher weighting. For effectively combining the evidences, we first classified the obtained evidences into two groups namely, overlapping and non-overlapping groups. Evidences are said to belong to the overlapping group if a minimum overlap of 70% exists between the two types of evidences. On the other hand, if the overlap is less than 70% then those evidences are referred to as non-overlapping.

In the case of overlapping evidences, starting and ending points of the combined evidence are determined as follows: If the starting point of an evidence detected using the proposed feature lies within three analysis frames (240 samples) of the

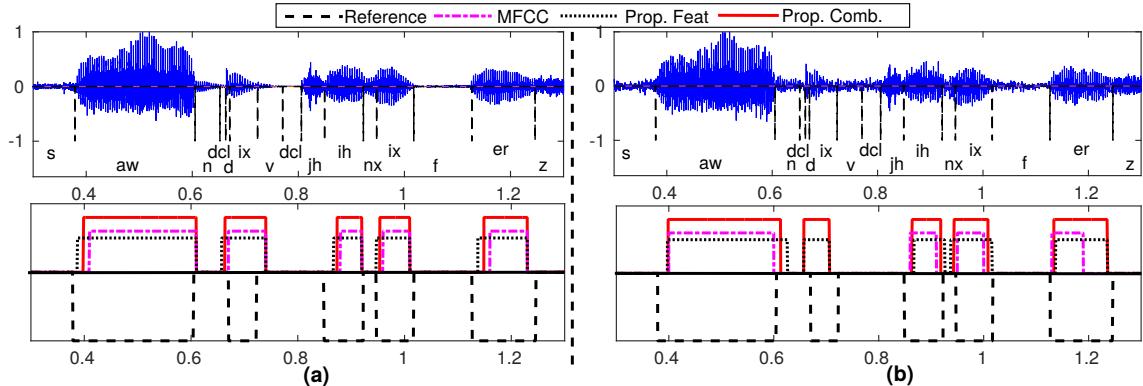


Fig. 2: Time-domain waveforms for a segment of speech along with the vowel evidences obtained using the MFCC features and the proposed features are shown in this illustration. The final evidences obtained using the proposed approach for combining the evidences are also shown. The ground truth obtained from the database are displayed as the reference markings for proper comparison. The efficacy of the proposed approach for combining the evidences is studied under (a) clean and (b) noisy condition when 5 dB babble noise is added to the speech signal. In each of the figures, the x-axis represents the time in seconds.

starting point of the evidence detected using MFCC, then the mean of those two locations is considered as the starting point of the final evidence. Else, the starting point of the vowel evidence detected by the proposed features is designated as the starting point of the final evidence. The end points in the case of overlapping evidences are also determined similarly. Those non-overlapping evidences that are a minimum of 100 ms in duration are identified and preserved in the final evidences without any modification. At the same time, we have treated those non-overlapping evidences that are less than 100 ms in duration as spurious detections and are eliminated.

A segment of speech data along with the evidences for the vowel regions obtained by using the MFCC and the features are shown in Figure 2. We have also plotted the final evidence obtained by combining the two evidences following the proposed approach to illustrate its effectiveness. The reference markings are also displayed for proper comparison. This analysis was performed under clean and noisy conditions i.e., when speech data is contaminated by 5 dB babble noise. The observed differences in the obtained evidences may probably be attributed to the fact that the proposed features and MFCC model the few frames near vowel transitions quite differently. On comparing the detected vowel evidences with the references, the proposed method of combining the evidences helps in detecting the vowel regions more accurately. This point is substantiated through the *IR* and *SR* values presented in Table II. Weighted combination of evidences following the proposed procedure is observed to be superior in clean as well as noisy conditions.

C. Comparing with existing techniques

For performance comparison, three existing VOP detection methods are explored in this study [2], [3], [8]. In [2], the employed features are Hilbert envelop of the linear prediction (LP) residual signal computed from the given speech signal, sum of the ten largest peaks in the short-term discrete Fourier transform (DFT) spectrum and the energy corresponding to 4-16 Hz components. This approach is termed as *COMB-EVI* in the rest of the paper. In the second VOP detection

TABLE II: Performances for the proposed and exist techniques for detecting vowels under clean and noisy test conditions.

SNR	Method	IR in %	SR in %		
			Semivowel	Nasal	Other
Clean speech					
	MFCC	84.76	10.93	1.80	4.32
	Prop. Feat	82.70	9.84	3.73	2.83
	Prop Comb	89.06	11.05	2.64	4.04
	COMB-EVI	73.07	9.54	1.92	18.79
	SE-GCI	66.81	10.72	1.48	3.68
	NLM-SE	67.93	12.47	1.40	7.24
Speech data degraded by Babble noise					
10 dB	MFCC	71.03	11.06	1.00	4.82
	Prop. Feat	75.46	10.02	3.15	4.38
	Prop. Comb	82.86	10.98	2.71	5.04
	COMB-EVI	70.02	9.79	2.23	25.85
	SE-GCI	63.64	11.14	1.98	21.65
	NLM-SE	68.95	11.75	1.77	10.73
5 dB	MFCC	52.14	11.13	0.50	5.01
	Prop. Feat	72.16	10.08	2.88	7.13
	Prop. Comb	75.00	10.95	2.71	7.41
	COMB-EVI	68.07	9.84	2.64	27.30
	SE-GCI	62.58	11.54	2.37	24.55
	NLM-SE	69.44	10.67	2.37	20.02
0 dB	MFCC	21.84	11.73	0.44	5.88
	Prop. Feat	60.03	10.08	2.78	7.63
	Prop. Comb	62.75	10.56	2.72	7.89
	COMB-EVI	67.68	10.19	2.92	29.83
	SE-GCI	61.78	11.82	2.83	29.80
	NLM-SE	67.09	10.18	3.19	25.44

approach [3], zero frequency filtering (ZFF) of speech signal was done to determine the glottal closure instants (GCIs). The short-term DFT magnitude spectrum was then computed

for the speech samples that are present in the 30% of a glottal cycle. Limiting the operation within the frequency band of 500-2500 Hz, the spectral energies for those regions was computed and used as the feature. In remaining of the paper, this approach is termed as *SPE-GCI*. Finally, proposed approach has also been compared with a recently reported approach for detecting vowels under noisy conditions [8]. In this method, an estimate of the speech signal at each time instant was obtained using non-local means (NLM) estimation. Then the cumulative sum of the short-term DFT spectrum was used as the front-end feature for detecting VOPs. In rest of the paper this approach is termed as *NLM-SE*.

The *IR* and *SR* values obtained by using COMB-EVI, SE-GCI and NLM-SE techniques under clean and noisy test conditions are enlisted in Table II. On comparing the *IR* and *SR* values, employing statistical models for detecting vowels is noted to be superior to the existing approaches especially under clean test conditions. At the same time, the proposed approach for combining the evidences outperforms all other schemes for detecting vowel regions. Even in terms of *SR*, the proposed technique for combining the evidences is superior to the existing methods. It is to note that, for 0 dB babble noise case, *IR* for the existing techniques is higher than that for the proposed approach. At the same time, *SR* values are also extremely high for the existing methods which is not desirable.

IV. CONCLUSION

A novel front-end feature extraction method has been proposed in this paper. The proposed features are used to develop a LSTM-based three-class classifier for detecting the vowel regions. Another three-class classifier is developed using the conventional MFCC features. In order to detect the vowel regions, frame-level alignments are generated for the given test utterance by forced-aligning it with respect to trained acoustic models. The proposed features result in better detection when speech-like noise is added to the test data. In addition to that, a novel and simple method for combining the evidences obtained using the MFCC and proposed features is also proposed to detect vowel regions with higher accuracy. By combining the evidences, the *IR* and *SR* values are noted to be much superior to those obtained by using the existing state-of-the-art approaches.

REFERENCES

- [1] D. J. Hermes, "Vowel onset detection," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866–873, February 1990.
- [2] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on audio, speech, and language processing*, vol. 17, no. 4, pp. 556–565, May 2009.
- [3] A. Vuppala, J. Yadav, S. Chakrabarti, and K. S. Rao, "Vowel onset point detection for low bit rate coded speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1894–1903, August 2012.
- [4] J. Yadav and K. S. Rao, "Detection of vowel offset point from speech signal," *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 299–302, April 2013.
- [5] K. N. Stevens, *Acoustic Phonetics*. The MIT Press Cambridge, Massachusetts, London, England, 2000.
- [6] K. Vuppala, K. S. Rao, and S. Chakrabarti, "Improved vowel onset point detection using epoch intervals," *AEU-International Journal of Electronics and Communications*, vol. 66, no. 8, pp. 697–700, August 2012.
- [7] S. R. M. Prasanna and G. Pradhan, "Significance of vowel-like regions for speaker verification under degraded condition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2552–2565, May 2011.
- [8] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Non-local estimation of speech signal for vowel onset point detection in varied environments," *Interspeech 2017*, pp. 429–433, August 2017.
- [9] S. Deb and S. Dandapat, "Emotion classification using segmentation of vowel-like and non-vowel-like regions," *IEEE Transactions on Affective Computing*, vol. 99, pp. 1–1, July 2017.
- [10] K. S. Rao and B. Yegnanarayana, "Duration modification using glottal closure instants and vowel onset points," *Speech Communication*, vol. 51, no. 12, pp. 1263–1269, December 2009.
- [11] A. K. Vuppala and K. S. Rao, "Vowel onset point detection for noisy speech using spectral energy at formant frequencies," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 229–235, June 2013.
- [12] A. Kumar, S. Shahnawazuddin, and G. Pradhan, "Exploring different acoustic modeling techniques for the detection of vowels in speech signal," in *Proc. National Conference on Communication*, March 2016, pp. 1–5.
- [13] ———, "Improvements in the detection of vowel onset and offset points in a speech sequence," *Circuits, Systems, and Signal Processing*, vol. 36, no. 6, pp. 2315–2340, June 2017.
- [14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.
- [15] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.
- [16] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Linguistic Data Consortium, December 1993, vol. 33.
- [17] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531–544, February 2014.
- [18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Workshop on automatic speech recognition and understanding*, December 2011.
- [19] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

Modelling Glottal Flow Derivative Signal for Detection of Replay Speech Samples

Jagabandhu Mishra *, Debadatta Pati † and S. R. Mahadeva Prasanna*

* Department of Electrical Engineering
Indian Institute of Technology Dharwad, Dharwad-580011, India

† Department of Electronics and Communication Engineering
National Institute of Technology Nagaland, Dimapur, Nagaland-797103, India
Email: 183081002@iitdh.ac.in, debapati2003@yahoo.com, prasanna@iitdh.ac.in

Abstract—It is a widely known fact that automatic speaker verification systems are quite vulnerable to replay speech. The present work deals with detecting replay speech by using the information available in glottal flow derivative (GFD) signal. In signal processing terms, the speech signal can be represented as the response of a vocal-tract system with excited by a excitation source in the form of *glottal flow*. The effect of record and replay devices distorted the spectral characteristics of the naturally uttered speech sample, resulting distortion in corresponding GFD signals. In this work the GFD signals are parameterized by using standard mel filters and Gaussian mixtures models are made for detection.

Although various methods are available, by correlation analysis it is observed that in the context of the present work the dynamic programming phase slope algorithm (DYPDA) method is relatively more effective in estimating the GFD signals. The experimental studies are made on ASVSpoof2017 database. The proposed glottal flow derivative mel frequency cepstral coefficients (GFDMFCC) feature provides 20.53% equal error rate (EER). This performance is comparatively poor than by speech and residual based features. It is mainly due to the absence of fine structure information in estimated GFD signal. However, in fusion with speech signal based constant-Q cepstral coefficients (CQCC) features, the GFDMFCC feature provides an improvement of 10.30% with reference to conventional residual feature. This shows the usefulness of modelling GFD signals for detection of replay signals.

I. INTRODUCTION

In Automatic speaker verification (ASV), the identity of a speaker is verified by the machine using the speaker specific clues available in the speech sample. Of late the ASV systems are widely used for various applications [1]. The major threat to ASV system is the risk of spoof attempts [2]. The most easily accessible and effective spoofing approach is *replay attacks*, where an impostor can use the pre-acquired speech samples of any target speaker. The current work deals with the detection of replay speech samples.

In general the replay sample can be detected by tracing the devices characteristics of the replay sample. The characteristics of the record and playback devices distort the nature of the original signal. For example, multiple quantization for recording and significant low-frequency attenuation by loud speaker for replay [3]. It is clearly visible from the

time domain representation of actual and corresponding replay speech samples shown in Fig. 1. But, such distortions are not easily distinguishable in spectral domain representations, also shown in Fig. 1. The reason may be the use of good quality record and playback devices that retain the spectral patterns of the original signal. This is also the reason for which the ASV systems are vulnerable to replay speech. It is conjecture that we may achieve more benefit in using the time domain features, such as excitation characteristic for detection of replay signals.

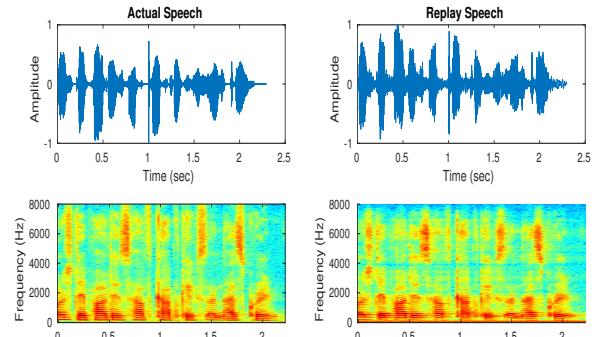


Fig. 1. Temporal and spectro-graphical representation of actual and corresponding replay samples.

The excitation is nearly a periodic air flow through glottis, called as glottal flow [4]. Due to the differentiation property of the lips, in uttered speech the presence of glottal flow is reflected by its derivative termed as glottal flow derivative (GFD) signal. Thus, the GFD signal can be treated as a representation of excitation signal. In our work we explore the GFD signal for detection of replay signals.

The paper is organized as follows: Section II presents a brief review on recent works on detection of replay signals and motivation for processing of GFD signals. With reference to the task objective Section III describes the computation and modelling methods of the GFD signal. In Section IV we discuss the experimental results and observations. Finally, we summarized the work with in Section V.

II. DETECTION OF REPLAY SIGNALS: A REVIEW

The INTERSPPECH 2017 organized a session on replay detection and has reported the effectiveness of several features. These include constant-Q cepstral coefficients (CQCC), Mel-frequency cepstral coefficients (MFCC), inverted Mel-frequency cepstral coefficients (IMFCC), subband spectral centroid magnitude coefficients (SCMC) and so others. The performance of these features are presented in Table I. Independently, the SCMC features provides the best performance of 11.49%. The excitation source based linear prediction (LP) residual features are also found to be effective in detection of replay signals. For example, the residual linear prediction cepstral coefficients (LPCCres) provides the detection performance of 27.61% [5]. The fusion of LP residual information with CQCC further improves the baseline detection performance by 10% [5]. These achievements show the importance of using source information. However, the LP residual is a parametric representation of the excitation signal, where as GFD is an estimation. Thus, GFD signal may contain relatively richer information regarding excitation component. It is therefore interesting to explore the GFD signal for detection of replay signal.

TABLE I
THE PERFORMANCE STATISTICS OF VARIOUS REPLAY ATTACK DETECTION FEATURES.

Features	Evaluation EER
CQCC [1]	24.77
MFCC [6]	27.12
IMFCC [6]	30.91
SCMC [6]	11.49
VESA-IFCC [7]	14.06
CQCC(6-8 kHz) [5]	17.31
LPCCres(6-8 kHz) [5]	27.61
CQCC+LPCCres [5]	13.95

The typical glottal flow signal and its derivative signals are shown in Fig. 2 [4]. The whole cycle is divided into three segments: *closed phase*, *open phase* and *return phase*. The closed phase corresponds to time interval the vocal folds are closed, and so no air flow. The open phase corresponds to the time interval during which vocal folds are either fully or partially open, and there is nonzero airflow. The return phase is related with completion of the speech production mechanism. It is defined as the time interval from the most negative value of the glottal flow derivative to the time of glottal closure. The return phase determines the amount of high-frequency energy present in the source and the speech signal. This component is especially affected by the response of the record and replay devices. For example, the multiple recordings introduce aliasing effect that affect the high frequency regions. The loudspeaker causes sharp attenuation in low frequency region causing disturbances in spectral flatness [3]. It can be observed from Fig. 2 that the rapidness of glottal activities are relatively more clear in GFD signal. This observation motivates us to model the GFD signal. In the following section we compute and model the GFD signal for detection of replay signals.

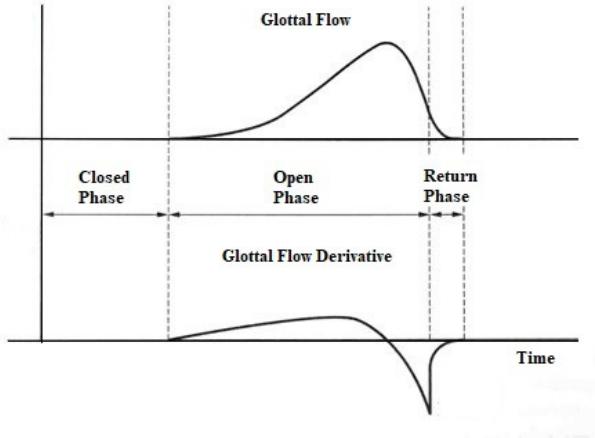


Fig. 2. Typical glottal flow and its derivative waveforms [4].

III. COMPUTATION AND MODELLING OF GFD SIGNAL

The glottal flow and its derivative are assumed to be consist of the components of coarse and fine structure [4], [8]. The coarse structure gives the general shape of the glottal cycle, where as the fine structure represents the ripples and aspiration. In literature several algorithms are available for estimation of GFD signal. The most contemporary methods include DYPSA [9], yet another GCI algorithm (YAGA), zero frequency resonator (ZFR) based method, speech event detection using the residual excitation and a mean-based signal (SE-DREAMS) and dynamic plosion index (DPI) algorithms [10], [11]. These algorithms equally outperform than other, although rely on different approaches. For example, the DYPSA and DPI approaches follow dynamic programming, where as the others follow smoothing process. Also, unlike others the DPI algorithm uses the residual signal. Thus these techniques may have different properties in terms of various parameters like accuracy, reliability and robustness. They are here evaluated in the context of task objective. For instance, how accurately the estimated GFD signal regenerate the original speech. If any deviation appear due to replay activity, then the corresponding GFD signal may regenerate different synthesized speech. Tracing such differences may be useful in detecting the replay signal.

The linear prediction coding (LPC) approach is employed for speech synthesis [12]. In terms of popular LP speech production model, the speech signal signal $s(n)$ can be expressed as,

$$s(n) = \hat{s}(n) + r(n) \quad (1)$$

and

$$\hat{s}(n) = \sum_{k=1}^p a_k s(n - k) \quad (2)$$

Where, with proper LP order the $\hat{s}(n)$ models the vocal-tract component of the actual speech signal in terms of LP

coefficients (LPCs) ‘ $a_k s$ ’. The error in the prediction $r(n)$, called as LP residual models the excitation component [13]. The LP residual signal is computed by using the following equation.

$$r(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (3)$$

The LP process synthesize the speech through inverse filtering the $r(n)$ signal with ‘ $a_k s$ ’ as filtering parameters [12]. The $r(n)$ signal is synonymous representation of the GFD signal. Various methods of glottal closing instant (GCI) computation is employed to compute the GFD signal using complex cepstrum decomposition method as mentioned in [14]. The GFD signals then analyzed using a correlation method and spectral distortion method to get the best estimation in context of replay detection. The synthesized speech and the correlation have computed using the equation. 4 and 5.

$$s_l(n) = Z^{-1} \left[\frac{Z[e_l(n)]}{1 - \sum_{k=1}^P a_k z^{-k}} \right] \quad (4)$$

$$Corr_l = \frac{1}{L} \left[\frac{\sum_{n=1}^L (s_r(n) - \overline{s_r(n)})(s_l(n) - \overline{s_l(n)})}{\sqrt{Var(s_r(n))}\sqrt{Var(s_l(n))}} \right] \quad (5)$$

where $s_l(n)$ is the synthesized speech, Z represents Z transform, L represents the length of the signal, l represents the index ($1 \leq l \leq 5$) for five different GFD signal and $e_l(n)$ is the GFD signals estimated using various methods. The correlation is computed between the synthesized speech using various GFD estimation techniques with a reference synthesized speech ($s_r(n)$) where the $e_l(n) = r(n)$. The method showing higher correlation is considered for analysis.

$$SD = \sqrt{\frac{2}{N} * \sum_k \left[10 \log_{10} |S_a(k)|^2 - 10 \log_{10} |S_r(k)|^2 \right]^2}, \text{ for } 0 \leq k < \frac{N}{2} \quad (6)$$

Table II shows the correlation values estimated from the speech samples of five different speakers for five different methods. On an average the DYPSA shows the highest correlation value of 0.720 followed by DPI with 0.665. These two methods provide almost equal correlation. For concrete clarification of choosing a method, we further compare these two methods with reference to our task objective. The GFD signals are estimated from the replay counterparts of the speech samples. The spectral distortion of actual and correspond replay speech samples are compared. The method with high spectral distortion is considered for experimental analysis. Table III shows spectral distortion values. The spectral distortion has computed using equation 6, where $S_a(k)$ and $S_r(k)$ are the estimated actual and replay GFD signals. It is clearly observed that the DYPSA approach provides relatively higher distortion value and thus considered for estimation of GFD signal. The estimated GFD signal of actual and replay signal using DYPSA and DPI method are depicted in Fig. 3. From the figure, we observed the oscillation (Zero crossing) in the replay signal is more and the epoch strengths are also less in reply signal in compared with actual signal.

TABLE II
THE CORRELATION BETWEEN THE LP BASED SYNTHESIZED SPEECH SIGNALS DERIVED FROM LP RESIDUAL AND ESTIMATED GFD SIGNAL BY USING FIVE DIFFERENT METHODS. THE PARAMETERS ARE COMPUTED BY USING SPEECH SAMPLES OF FIVE DIFFERENT SPEAKERS.

Speakers	Correlation coefficient				
	SEDREAMS	ZFR	DYPSA	YAGA	DPI
Spk1	0.104	-0.083	0.685	0.237	0.696
Spk2	0.480	0.192	0.782	0.108	0.691
Spk3	0.021	0.002	0.727	0.090	0.652
Spk4	0.135	0.0566	0.708	0.096	0.681
Spk5	0.218	-0.1338	0.696	0.232	0.606
Average	0.191	0.006	0.720	0.153	0.665

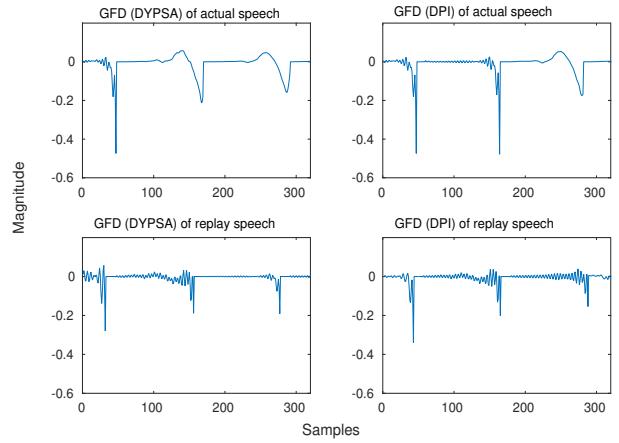


Fig. 3. Estimated GFD signal of actual and corresponding replay signal.

In modelling the GFD signal we use the advantage of mel-scale distribution (i.e tightly spaced from 0-500 Hz) and process the GFD signal through mel-filters and then compute the cepstral coefficients. In the region from 0-500 Hz the excitation source information is predominantly reflect in spectrum. We compute the cepstral features called as GFD mel-filter cepstral coefficients (GFDMFCC) using the standard cepstral analysis. The first thirteen coefficients including zeroth-one are used as representative features. The results of the performed experiments and the followed observations are presented in following Section IV.

TABLE III
SPECTRAL DISTORTION (SD) BETWEEN ACTUAL AND REPLAY GFD SIGNALS ESTIMATED BY DYPSA AND DPI METHODS. THE PARAMETERS ARE COMPUTED BY USING SPEECH SAMPLES OF FIVE DIFFERENT SPEAKERS.

Speakers	Spectral Distortions	
	DYPSA	DPI
Spk1	13.54	14.53
Spk2	13.27	13.06
Spk3	16.82	16.52
Spk4	14.35	13.42
Spk5	14.72	12.83
Average	14.54	14.07

TABLE IV
PERFORMANCE COMPARISON OF THE GFDMFCC FEATURE FOR REPLAY SIGNALS DETECTION TASK.

Features	Evaluation EER
GFDMFCC	20.53
RMFCC	14.75
CQCC	15.12
CQCC+RMFCC	10.29
CQCC+GFDMFCC	9.23

IV. EXPERIMENTAL STUDY

A. Experimental Setup

The experimental studies are made with ASVspoof2017 database [1], that contains actual and replay speech signals collected at 16 kHz. The GFD signals are estimated by using DYPSA approach and LP residual signals by using Equation 3. As suggested in [15], the LP order is chosen as 20. The GFDMFCC and residual mel-frequency cepstral coefficients (RMFCC) features are computed for every 20 mili-seconds with a shift of 10 mili-seconds. The actual and replay models are build using Gaussian mixtures model(GMM) classifier [6]. The actual and replay models has been build using training and development set data. The final performance of the system is evaluated with evaluation set. In the evaluation process, the ASVspoof2017 protocol is followed, where the trials are compared with actual and replay models. Finally the log likelihood ratio (LLR) are computed using Equation 7 [5]. The performance of the systems are shown in terms of EER and using detection error trade-off (DET) curves.

$$LLR = \log(L_{actual}) - \log(L_{replay}) \quad (7)$$

Where L_{actual} and L_{replay} are the LLRs computed using test utterance compared with actual and replay GMM models, respectively. Obtained results and observations are discussed below.

B. Experimental Results and Discussion

The results are given in Table IV, and corresponding DET curves are shown in Fig. 4. The proposed GFDMFCC feature provides the EER of 20.53%. In comparison the LP residual based feature provides an EER of 14.75%. On the other hand the speech signal based CQCC feature provides an EER of 15.12%. The GFDMFCC individual provides the poorest performance. The reason may be due to the presence of only coarse structure information in GFD signals. However, it is interesting to notice that as compared to RMFCC, the joint use of CQCC and GFDMFCC provides an improvement of 10.30%, indicating the significance of the later.

V. CONCLUSION

Spoofing by replay speech samples to ASV systems is a serious threat to their reliability. In this work we examine and demonstrate the potential of GFD signals for detection of replay signals. First, we investigate various GFD signal estimation methods and select the best possible approach,

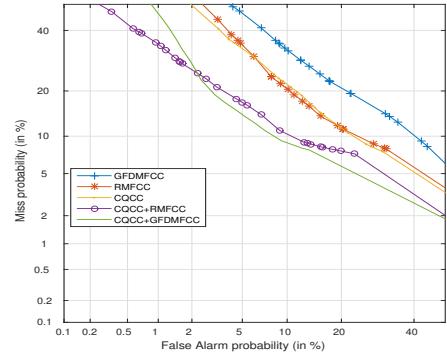


Fig. 4. DET curves showing the performance of replay attacks detection task with GFDMFCC, RMFCC, CQCC features and its fusions

particular in the context of replay detection tasks. In that sense the DYPSA approach of estimating GFD signal is found to be relatively more effective. With that representation mel-warped based cepstral parameterization is employed to model the GFD signal. In GMM based replay detection system the purposed GFDMFCC feature provides 20.53% with ASVspoof2017 database. The reason may be due to the absence of fine structure information. In comparison to excitation source information based residual feature, the joint use of proposed GFDMFCC and speech signal based constant-Q cepstral coefficients (CQCC) provides an improvement of 10.30%. The future plan is to use both course and fine structure information and estimate the GFD signal, then investigate its potential for detection of replay samples.

REFERENCES

- [1] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge:assessing the limits of replay spoofing attack detection," in *INTERSPEECH*, 2017.
- [2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 135–153, 2015.
- [3] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Lecture Notes in Computer Science. Springer*, pp. 274–285, 2011.
- [4] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modelling of glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech and Audio Process.*, vol. 7, pp. 569–586, Sep. 1999.
- [5] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features," in *INTERSPEECH*, 2017.
- [6] R. Font, J. M. Espin, and M. J. Cano, "Experimental analysis of features for replay attack detection results on the asvspoof 2017 challenge," in *INTERSPEECH*, 2017.
- [7] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length teager energy separation based instantaneous frequency features for replay detection," in *INTERSPEECH*, 2017.
- [8] T. V. Ananthapadmanabha and G. Fant, "Calculation of true glottal flow and its components," *Speech Commun.*, vol. 1, pp. 167–184, Dec. 1982.
- [9] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the dyspa algorithm," 2007.
- [10] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, pp. 994–1006, March 2012.

- [11] A. Prathosh, T. Ananthapadmanabha, and A. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2471–2480, 2013.
- [12] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.
- [13] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Commun.*, vol. 48, pp. 1243–1261, Jun. 2006.
- [14] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Communication*, vol. 53, no. 6, pp. 855–866, 2011.
- [15] J. Mishra, M. Singh, and D. Pati, "Lp residual features to counter replay attacks," in *Signals and Systems (ICSigSys), 2018 International Conference on*, pp. 261–266, IEEE, 2018.

Comparison of low-dimension speech segment embeddings: Application to speaker diarization

Srikanth Raj Chetupalli

Dept. of ECE

*Indian Institute of Science
Bangalore, India*

Sreenivas Thippur V.

Dept. of ECE

*Indian Institute of Science
Bangalore, India*

Anand Gopalakrishnan

Dept. of EEE

*National Institute of Technology, Surathkal
Karnataka, India*

Abstract—Segment clustering is a crucial step in unsupervised speaker diarization. Bottom-up approaches, such as, hierarchical agglomerative clustering technique are used traditionally for segment clustering. In this paper, we consider the top-down approach to clustering, in which a speaker sensitive, low-dimensional representation of segments (speaker space) is obtained first, followed by Gaussian mixture model (GMM) based clustering. We explore three methods of obtaining the low dimension segment representation: (i) multi-dimensional scaling (MDS) based on segment to segment stochastic distances; (ii) traditional principal component analysis (PCA), and (iii) factor analysis (i-vectors), of GMM mean super-vectors. We found that, MDS based embeddings result in better representation and hence result in better diarization performance compared to PCA and even i-vector embeddings.

I. INTRODUCTION

Speaker diarization is a speech segmentation task involving unsupervised identification of speakers, addressing the problem of ‘who spoke when?’ in a conversation. Speaker diarization is a major component of speaker indexed information retrieval with applications to indexing of telephone conversations, meeting recordings, and broadcast TV shows etc. Several approaches have been proposed in the literature to address the speaker diarization problem; a review of this is available in [1], [2], [3].

Most of the approaches to speaker diarization consist of two steps. In the first step, speaker homogeneous segments are identified, and in the second step, segments are clustered to obtain speaker labels. Bayesian information criterion (BIC) based techniques are traditionally used to obtain the initial segmentation [4], and then agglomerative clustering [5] is used to identify and associate the speaker segments. However, agglomerative clustering method is prone to errors in the identification of short segments at the leaf nodes and hence affect the performance of speaker-segment association. To overcome this, top-down clustering based approaches have also been considered [6]. In [6], the spectral clustering approach is considered using KL divergence distance measure between segment models. Motivated by the success of factor analysis in speaker identification tasks [7], [8], i-vector based approaches are applied to speaker diarization also [9], [10]. In [9], a variational Bayes approach is developed to classify segments in the speaker factor (i-vector) space, in which

the initial speaker homogeneous segments are assumed to be known. However, they use initial segmentation using BIC based criterion and hence they are prone to errors. Another approach based on uniform overlapped segments and i-vector based segment clustering is considered in [10]. Uniform segments with successive overlap are drawn from the speech conversation, and i-vectors are estimated using pre-trained GMM-UBM (universal background model) [11] and factor analysis matrix. The segment i-vectors are then projected into a low-dimensional space using PCA followed by GMM based clustering. The soft segmentation obtained after GMM based clustering is used to build individual speaker models and then update the segment boundaries using Viterbi re-segmentation. The method is performed in an iterative manner, with the uniform segments obtained as estimated segment boundaries from the previous iteration. Pre-trained UBM and the total variability matrix for i-vector estimation are used in [10]. However, speaker diarization being a single conversation specific task, the intra-conversation variability of speakers causes the dominant distortion and pre-trained models using a training corpus (different from the test corpus) may not generalize well to the specific conversation under test.

In factor analysis and spectral clustering based approaches discussed above, the segments are first modeled as GMMs and embedded into a “speaker space” followed by clustering in the speaker space [6], [10]. For a properly chosen distance measure in the probability distribution space, segments from same speaker should have smaller distances and segments from different speakers should have higher distances, for better clustering leading to better segmentation even with overlapping speaker data. Motivated by this, we propose a low-dimensional representation of segment models using distance matrix based embedding, i.e., we seek a realization of points in a low-dimensional space such that the pair-wise distances between the segment representations equal the corresponding distance between the segment models. This can be achieved using multi-dimensional scaling (MDS) [12]. The segment models are MAP adapted from the UBM and symmetric KL divergence is used as the distance measure for the distance matrix based embedding. Mean only MAP adaptation is typically used, but we explore adaptation of the weight and covariance parameters along with the mean to estimate the segment models. We show that adaptation of all UBM

parameters does improve the contrast between segments from different speakers which results in better embedding compared to mean only adaptation. We compare the distance matrix based approach with the traditional i-vector analysis, and PCA projection of GMM super-vectors and show that the distance matrix based approach performs better both qualitatively as well as in terms of diarization error rate (DER).

II. LOW DIMENSIONAL SEGMENT REPRESENTATION

Let $\mathcal{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ be the sequence of feature vectors extracted from the given conversation, and let $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K\}$ be K successive segments taken from \mathcal{X} . For the start and end points n_k^l, n_k^u with $1 \leq n_k^l < n_k^u \leq N$, we have $\mathbf{X}_k = [\mathbf{x}_{n_k^l}, \mathbf{x}_{n_k^l+1}, \dots, \mathbf{x}_{n_k^u}]$. GMM-UBM framework is used to model each segment \mathbf{X}_k and an M mixture GMM, estimated using all the feature vectors of \mathcal{X} is used as the UBM.

$$p(\mathbf{x}|\Theta) = \sum_{m=1}^M \alpha_m \mathbb{P}(\mathbf{x}|\theta_m), \quad (1)$$

where $\Theta = \{\alpha_m, \theta_m, 1 \leq m \leq M\}$ is the set of parameters of GMM, and $\theta_m = \{\mu_m, \Sigma_m\}$ denotes the mean and variance of the individual Gaussian components. Then the density function of k^{th} segment is obtained via MAP adaptation using the feature vectors of the segment \mathbf{X}_k ,

$$\mathbb{P}(\mathbf{x}|\Theta_k) = \sum_{m=1}^M \alpha_{km} \mathbb{P}(\mathbf{x}|\theta_{km}). \quad (2)$$

by updating all the parameters α_{km} and θ_{km} .

\mathbf{x}_n is 12-dimensional mel-frequency cepstral coefficients (MFCC) (without energy), extracted at 100 Hz rate using 25 ms windows and number of mixtures in UBM is chosen as $M = 32$. We note that a higher number of mixture components is typically used for speaker ID task; but, for unsupervised diarization a smaller number of mixture components is found to be sufficient. Also, in speaker recognition applications, mean only adaptation is often used and found to be sufficient; however, we explore here MAP adaptation of all the parameters of UBM to model the segments.

From the estimated parametric segment model $\mathbb{P}(\mathbf{x}|\Theta_k)$, we would like to represent it in a low-dimensional space such that unnecessary variability due to speech text or recording medium is reduced. Thus a good low-dimensional representation brings out the contrast between the different speaker segments and enables better clustering of the segment models in the “speaker space”. We consider distance matrix based low-dimensional representation to achieve this. Let d_{ij} be a measure of distance between two segment models $\mathbb{P}(\mathbf{x}|\Theta_i)$, $\mathbb{P}(\mathbf{x}|\Theta_j)$ for $\{i, j\} \in [1, 2, \dots, K]$, and the matrix of pairwise distances between segment models be $\mathbf{D} \in \mathcal{R}^{K \times K}$, where $[\mathbf{D}]_{ij} = d_{ij}^2$. The segment embedding problem is defined as that of finding a set of points $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K\}$ in a $r < K$ -dimensional space, such that the pair-wise Euclidean distance

matrix of the mapped points in the r -space $\approx \mathbf{D}$, i.e.,

$$\begin{aligned} &\text{find } \mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_K \\ &\text{s.t } d_{ij} \approx \|\mathbf{c}_i - \mathbf{c}_j\|_2, \forall \{i, j\} \in [1, \dots, K] \end{aligned} \quad (3)$$

The problem can be solved using the classical “multi-dimensional scaling” approach [12]. The computational steps involved in the solution of (3) are outlined in Alg. 1.

Algorithm 1 Multi-Dimensional Scaling (MDS)

- 1: Inputs: $d_{ij} \forall i, j \in [1, 2, \dots, K]$, target dimension r .
- 2: Form squared distance matrix $\mathbf{D} \in \mathcal{R}^{M \times M}$: $[\mathbf{D}]_{ij} = d_{ij}^2$.
- 3: Perform centering of the squared distance measurements

$$\mathbf{G} = -\frac{1}{2} \mathbf{J} \mathbf{D} \mathbf{J}^T$$

where the centering matrix $\mathbf{J} = \mathbf{I} - \frac{1}{K} \mathbf{1} \mathbf{1}^T$, \mathbf{I} is an identity matrix, and $\mathbf{1}$ is a vector of all 1’s.

- 4: Compute the eigen vectors corresponding to the r -largest eigen values of \mathbf{G} ,
- 5: Obtain the low-dimensional representation as

$$\mathbf{C} = [\mathbf{c}_1 \dots \mathbf{c}_K] = \Lambda^{\frac{1}{2}} \mathbf{V}^T.$$

For $d_{ij} = \text{symmetric KL divergence}$, it is defined as,

$$d_{ij} = D_{sKL}(i, j) = D_{KL}(i, j) + D_{KL}(j, i), \quad (4)$$

$$\text{where } D_{KL}(i, j) = \int \mathbb{P}(\mathbf{x}|\theta_i) \log \left[\frac{\mathbb{P}(\mathbf{x}|\theta_i)}{\mathbb{P}(\mathbf{x}|\theta_j)} \right] d\mathbf{x}.$$

The symmetric $D_{sKL}(i, j)$ is used because KL divergence D_{KL} is asymmetric which is not suitable as a distance matrix. Also KL divergence is preferred because we have adapted all the UBM parameters. However, for GMMs, D_{KL} between two models can not be computed in a closed form, hence sampling based method is used; we use the voicebox [13] implementation, which computes the KL divergence using variational approximation [14].

The distance matrix based low dimensional representations are compared with the traditional GMM super-vector and i-vector based approaches. For the GMM super-vector approach low dimensional representation is obtained using principal component analysis (PCA) of the GMM super-vectors of all segment models, also referred to as eigen-voice space. Let $\{\mathbf{m}_1, \dots, \mathbf{m}_K\}$ denote the segmental GMM super-vectors (concatenation of all the mean vectors of Gaussian components of the model) of the K number of segments. PCA embedding is obtained as the projection of segmental GMM super-vectors onto the first r eigen vectors of the covariance matrix $\mathbf{R} = \frac{1}{K} \sum_{k=1}^K (\mathbf{m}_k - \mathbf{m}_0)(\mathbf{m}_k - \mathbf{m}_0)^T$, where \mathbf{m}_0 is the mean of all GMM super-vectors. The low dimension vectors

$$\mathbf{c}_k^p = \mathbf{U}_r^T (\mathbf{m}_k - \mathbf{m}_0), \quad (5)$$

where the matrix \mathbf{U}_r contains the first r eigen vectors of \mathbf{R} .

In the case of i-vector factor analysis, the GMM super-vectors are modeled as,

$$\mathbf{m}_k = \mathbf{m}_0 + \mathbf{T}\mathbf{c}_k^i, \quad k = 1, \dots, K \quad (6)$$

where \mathbf{T} is the factor matrix, and \mathbf{c}_k^i is the latent factor loading vector. \mathbf{c}_k^i is assumed to be a zero-mean unit-variance Gaussian random vector, and the maximum likelihood estimate of \mathbf{c}_k^i is used as the speaker representation (i-vector). In this paper, i-vectors are extracted using the method proposed in [8].

Thus, we are comparing three different low-dimensional speech segment representations, viz., \mathbf{c}_k : MDS based on D_{sKL} separability, \mathbf{c}_k^p : PCA of direct GMM super vector separability, and \mathbf{c}_k^i : i-vector derived separability.

III. SPEAKER DIARIZATION

The low dimensional segment representation methods described in the previous section are evaluated in the diarization scheme shown in Fig. 1. In the first step, MFCC features are extracted, and a 32 component GMM is trained using the whole conversation of about 180 s, which acts as the UBM $\mathbb{P}(\mathbf{x}|\Theta)$, to obtain the individual segment models. A simple voice activity detection using frame-energy thresholding is performed to exclude non-speech frames in the UBM. In the second step, uniform segments are formed using a sliding window of 2 sec duration, and successive shift of 400 ms. Unlike traditional approaches to speaker diarization, which use change point detection algorithm followed by clustering, we use clustering of the models derived from the uniform segments $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_K\}$. This has the advantage of better representing the speaker-ID taking care of the text variability in speech, avoiding a poorly performing segment contrast distance measure. Further, a low-dimensional representation of the segments, $\{\mathbf{c}_1, \dots, \mathbf{c}_K\}$, is obtained using the embedding methods described.

The clustering of embedded segment representation vectors leads to a second level Gaussian mixture modeling absorbing the intra conversation speaker variability. Here, we assume that the number of mixtures is equal to the number of speakers in the conversation and that it is known a-priori, since the present goal is only to evaluate different segment representations. However, in a practical system, the number of speakers also needs to be estimated, which can be done using variational Bayes GMM clustering [15], [10]. The speaker space Gaussian mixture posterior measure $\gamma(z_{ks}) \triangleq \mathbb{P}(z_{ks} = 1|\mathbf{c}_k)$ estimated in the iterations of the EM algorithm is the a-posteriori probability of k^{th} segment belonging to s^{th} speaker. Each frame posterior measure $\gamma(z_{ns}) \triangleq \mathbb{P}(z_{ns} = 1|\mathbf{x}_n)$, is assigned as the corresponding segment posterior probability,

$$\gamma(z_{ns}) = \gamma(z_{ks}), \quad \forall n_k^l \leq n \leq n_k^u. \quad (7)$$

Since the successive segments $\{\mathbf{X}_k\}$ have an overlap, the feature vector at instant n will be part of more than one successive segments (five for 2 s segments with 400 ms shift); hence, an average of the frame posteriors from successive segments is used as the estimate $\bar{\gamma}(z_{ns})$. From the averaged frame posterior $\bar{\gamma}(z_{ns})$, frame wise speaker label (of cluster indices)

is obtained using max-rule; and the segment boundaries are formed, for a tentative speaker activity graph.

Segment boundaries are further improved based on a GMM model $\mathbb{P}(\mathbf{x}|\Theta^s)$ obtained via MAP adaptation of UBM using the speaker segments determined in the previous step. The ML estimation approach [16] is then used for speaker spotting using the adapted speaker models. This can help refine boundary of segments constituting more than one speaker (however, it is possible to use other types of decoding such as Viterbi decoding using the individual speaker models). Thus, for a segment \mathbf{X}_k , speaker spotting problem is posed as,

$$\underset{\{\alpha_s\}}{\text{maximize}} \mathbb{P}(\mathbf{X}_k) \triangleq \sum_{s=1}^S \alpha_s \mathbb{P}(\mathbf{X}_k|\Theta^s) \quad (8)$$

The estimated α_s is interpreted as the posterior probability of speaker s , and the speaker label is obtained using max rule, $s^* = \arg \max \{\alpha_s\}$. From the segment-wise posterior speaker probabilities, frame-wise segmentation is obtained via hard thresholding similar to the GMM posteriors in the previous step.

IV. EXPERIMENTS AND RESULTS

Simulated conversation sequences are generated by concatenating speech from different speakers for evaluation and testing the embedding schemes and also choose optimum model parameters. This will also provide for harder evaluation with respect to number of speakers in a conversation, their random duration and text. A conversation having S number of speakers and $K = 40$ number of segments is generated as follows: (i) a random subset of S speakers is sampled from a pool of 12 speakers (6 male, 6 female), (ii) for each segment, a random speaker is selected from the chosen subset of 12 speakers and a continuous speech segment of duration 2 – 6 sec is clipped from the speech recording, and (iii) speech segments thus clipped are concatenated to generate the synthetic conversation of known segment boundaries. While concatenating, we made sure that successive segments belong to different speakers. Average duration of concatenated conversations is ≈ 180 seconds. The speech recordings are taken from [17], and they correspond to news bulletins broadcast in Hindi language. They have a lot of variability due to recording devices, different news readers, different dates of recording, etc. The original speech recordings are in MP3 format and with different sampling rates and channel conditions. We re-sampled all recordings to 8 KHz and converted to raw samples before concatenating to generate the synthetic conversation files.

A. Performance of segment embeddings

We now compare MDS, PCA and i-vector embeddings. For MDS embedding, we compare three alternatives of mean only adaptation (MDS-m), mean and variance (MDS-mv), and mean, variance and weight (MDS-mvw) adaptation methods. Fig. 2 shows an illustration of segment embedding in $r = 2$ dimensional space using a concatenated example conversation of $S = 3$ speakers. Segments are formed using a sliding

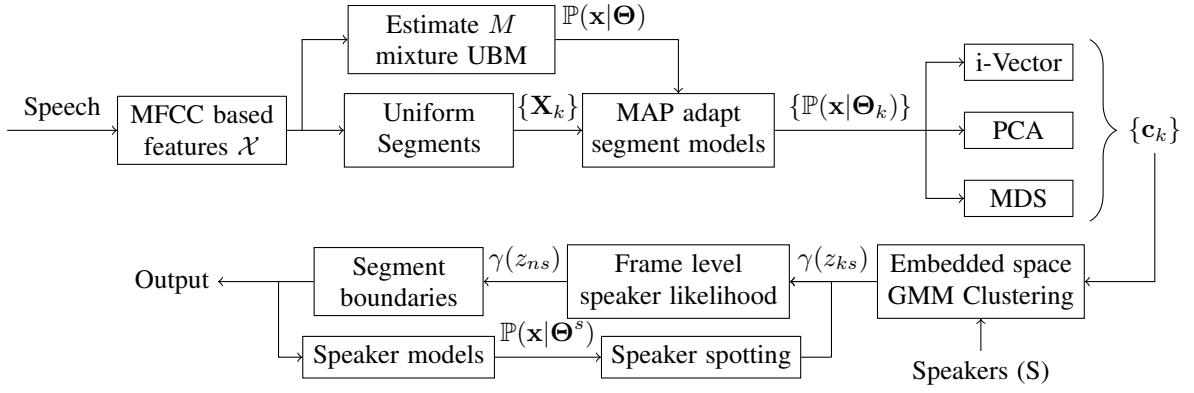


Fig. 1. Block diagram of unsupervised diarization

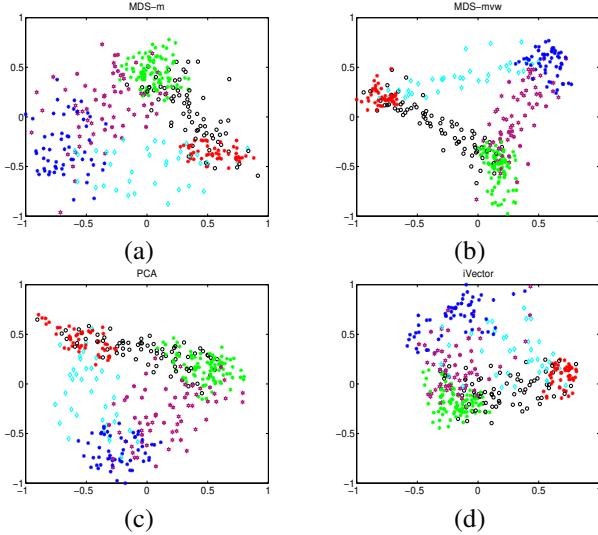


Fig. 2. Illustration of segment embedding for $r = 2$: using MDS method, where segment models obtained by the adaptation of (a) mean, (b) mean variance and weights; (c) PCA of mean super-vectors, and (d) i-vectors. Points in red, green and blue (speakers 1,2 and 3 respectively) correspond to speaker homogeneous segments, and other points in cyan, magenta and black correspond to mixed speaker segments.

window of 2 sec and a shift of 400 ms. We observe that the segment space is in general a continuum, for all the embedding methods; segments corresponding to same speaker (speaker homogeneous) (points in blue, green and red) get clustered together, and the segments comprising two speakers (speaker inhomogeneous) lie in the region between the clusters corresponding to the individual speakers. For example, the points in magenta correspond to segments with partial data from speakers 2,3 (blue and green), and lie in the region between the clusters formed by green (speaker 2) and blue points (speaker 3). A desirable property of segment embedding is that, with-in cluster scatter is less and the inter-cluster separation is more.

We measure the performance of each embedding method using the Fisher discriminant (FD) measure in the embedded space, defined as,

$$J = \text{tr}(\mathbf{S}_w^{-1} \mathbf{S}_b) \quad (9)$$

where \mathbf{S}_w and \mathbf{S}_b are the with-in-class and between-class scatter matrices respectively [15]. Segments are extracted uniformly with a sliding window, however only speaker homogeneous segments (obtained using speaker boundaries from ground truth labels) are used for the computation of Fisher measure. Fig. 3(a,b) show Fisher measure, averaged over 100 conversations each of ≈ 3 mins and 3 – 5 speakers. As a function of the embedding dimension r , the FD measure increases with r , as expected. MDS with KL divergence measure (obtained after adapting all parameters of the GMM) performs the best, distinctly better than other embeddings. It is clear that $\text{MDS-mvw} > \text{MDS-mv} > \text{MDS-m}$, clearly indicating that adapting the weights of UBM mixture components is important for speaker separability. The i-vectors perform poorest for small number of embedding dimensions. We found that PCA performs better than i-vectors and also MDS-m.

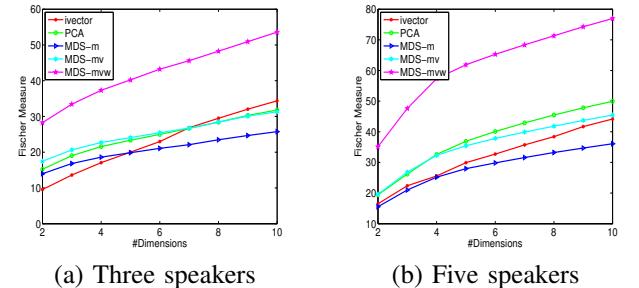


Fig. 3. Fisher measure as a function of number of embedding dimensions r for different embedding methods.

Uniform segmentation without known speaker boundaries does result in speaker inhomogeneous segments, and we have observed that factor matrix learned in the presence of such segments, result in a poor segment embedding compared to PCA and other MDS based embeddings. Thus MDS-mvw retains cluster separation even in the presence of speaker inhomogeneous segments.

B. Diarization performance

Fig. 4 shows the diarization performance in terms of diarization error rate (DER), and frame-wise F-Measure for different

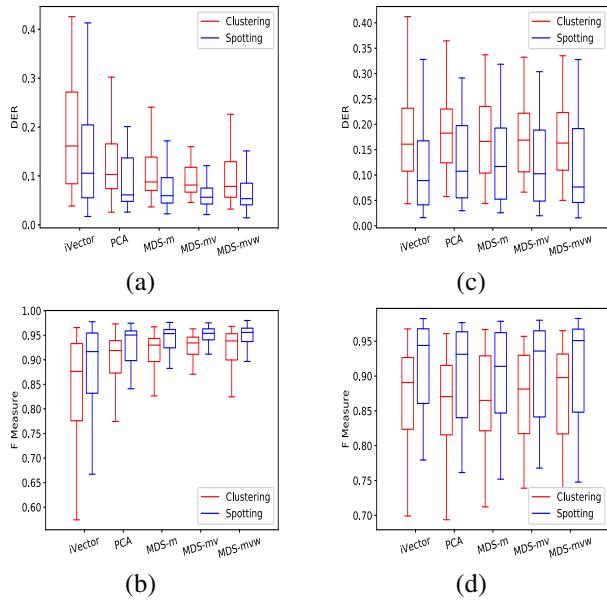


Fig. 4. Diarization performance for $r = 3$ dimensional embedding. The two columns show results for $S = 3$ and $S = 5$ speaker conversations.

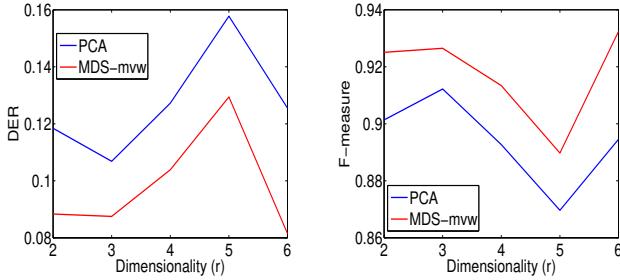


Fig. 5. DER performance as a function of the dimensionality of embedding space, for two embedding methods PCA and MDS-mvw. Number of speakers in the conversation is $S = 3$.

embeddings in $r = 3$ space. We see that, for three speaker conversations, MDS based embeddings perform better than PCA or i-vector based approaches. However, for conversations with five speakers, the performance is similar for all the methods compared. Here also in terms of DER, MDS-mvw is better than MDS-mv which is better than MDS-m. In all the cases, the “supervised spotting” stage after the GMM based clustering in the segment embedding space is found to improve the performance measures.

The performance as a function of the number of embedded dimensions is shown in Fig. 5. We see that best performance is obtained for $r = 3$ embedding, and compared to PCA, MDS-mvw embedding is found to perform better in-terms of both DER and F-measure. The plot shows that increasing the number of dimensions does not necessarily improve the diarization performance, and may also degrade the performance. This is because increasing the dimensionality of embedding space may create spurious clusters affecting the performance of the GMM clustering step and degrade the diarization performance.

V. CONCLUSIONS

Lower dimensional embedding is a means to reduce data variability due to unimportant aspects of the information in the signal. We have found MDS based embedding of speech segments with segment models computed using adaptation of all the parameters of the UBM-GMM, leading to good representation for speaker segregation and diarization. Also, a low dimension of $r < 5$ provides satisfactory results. The proposed low-dimensional representation is used for speaker diarization starting with uniform initial segmentation, and we found that the MDS based embeddings using model space distance perform significantly better than the traditional PCA and i-vector based methods. It may be useful to investigate other types of model space embeddings and its use for separating different types of speech information.

REFERENCES

- [1] S. E. Tranter and D. A. Reynolds, “An overview of automatic speaker diarization systems,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sept 2006.
- [2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.
- [3] M. Moattar and M. Homayounpour, “A review on speaker diarization systems and approaches,” *Speech Communication*, vol. 54, no. 10, pp. 1065 – 1103, 2012.
- [4] A. Tritschler and R. A. Gopinath, “Improved speaker segmentation and segments clustering using the bayesian information criterion,” in *EUROSPEECH*, 1999.
- [5] D. A. Reynolds and P. Torres-carrasquillo, “The mit lincoln laboratory rt-04f diarization systems: Applications to broadcast audio and telephone conversations,” in *in Proc. Fall 2004 Rich Transcription Workshop (RT-04), Palisades*, 2004.
- [6] H. Ning, M. Liu, H. Tang, and T. Huang, “A spectral clustering approach to speaker diarization,” in *Proc. ICSLP*, 2006.
- [7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.
- [8] P. Kenny, “A small footprint i-vector extractor,” in *Proc. ISCA Odyssey, The Speaker and Language Recognition Workshop, Singapore*, 2012, pp. 1–6.
- [9] P. Kenny, D. Reynolds, and F. Castaldo, “Diarization of telephone conversations using factor analysis,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 6, pp. 1059–1070, Dec 2010.
- [10] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, “Unsupervised methods for speaker diarization: An integrated and iterative approach,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2015–2028, Oct 2013.
- [11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.
- [12] W. S. Torgerson, “Multidimensional scaling: I. theory and method,” *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.
- [13] M. Brookes. (2003 (accessed Oct 11, 2017)) Voicebox: Speech processing toolbox for matlab. <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>.
- [14] J. R. Hershey and P. A. Olsen, “Approximating the kullback leibler divergence between gaussian mixture models,” in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, April 2007, pp. IV–317–IV–320.
- [15] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
- [16] H. Sundar, T. V. Sreenivas, and W. Kellermann, “Identification of active sources in single-channel convolutive mixtures using known source models,” *IEEE Signal Processing Letters*, vol. 20, no. 2, pp. 153–156, Feb 2013.
- [17] News on air. <http://newsonair.com/>.

Improved Epoch Extraction From Speech Signals Using Wavelet Synchrosqueezed Transform

D. Govind, S. Lakshmi Priya, S. Akarsh, B. Ganga Gowri, K. P. Soman

Center for Computational Engineering and Networking (CEN)

Amrita School of Engineering, Coimbatore

Amrita Vishwa Vidyapeetham, India

Email:d_govind@cb.amrita.edu, kp_soman@amrita.edu, {lakshmipriyashasi,akarshsoman,gangab.90}@gmail.com

Abstract—Synchrosqueezed wavelet transform (WSST) is an effective tool in tracking instantaneous frequency of a given signal. The objective of the present work is to propose a WSST based method for accurate epoch estimation from speech. Epochs in speech represent the instants where the excitation to the vocal tract is maximum and instantaneous F_0 contour is derived from epoch locations. The proposed hypothesis in this paper is that the signal reconstructed by discarding higher frequency modes (above the mean F_0) in the WSST transformed time frequency domain observed to predominantly represent source characteristics. The presence of the source characteristics in the modified WSST reconstructed signal is validated by the improved identification accuracy obtained for the epochs estimated from clean speech utterances of CMU-Arctic database. To further demonstrate the effectiveness of the WSST in improving the overall epoch estimation performance, a WSST modified zero frequency filtering (ZFF) of speech, which is one of the simple and effective tools for epoch extraction, is proposed. The sharp instantaneous frequency representation by WSST also found to be effective in estimating epochs emotion utterances where rapid pitch variations are present. The improved epoch estimation performance from emotive utterances are confirmed by validating on the German emotion speech corpus(EmoDb).

I. INTRODUCTION

Epochs represent locations where glottal closure instants occur in speech which also indicate the time instants at which excitations to the vocal tract are maximum. Epoch locations generally are perceptually relevant and high SNR regions and hence many applications such as speech enhancement, prosody modification, etc. are devised by using the knowledge of epochs. Due to the time varying nature of vocal tract response, determining the accurate locations of epochs and their strength is a challenging task. There exist many algorithms proposed for the estimation of epochs accurately and efficiently from speech signals.

Dynamic programming projected phase slope Algorithm (DYPFA) [1], group delay approach [2], Hilbert envelope [3], zero frequency filtering (ZFF) of speech [4], Speech Event Detection using the Residual Excitation and a mean-based signal (SEDREAMS) [5], dynamic plosion index based integrated linear prediction residual (ILPR) [6] and single frequency filtering (SFF) [7], [8] are the popular conventional epoch extraction methods from clean speech signals. The ZFF of speech and SEDREAMS are the earliest methods proposed by estimating the variations of the signal around the mono frequency components. In ZFF of speech, the variation of

signal around the zero frequency component is estimated and is found to be varying according to discontinuities due to epochs. In SEDREAMS, the regions around a mean based signal is estimated and instants of glottal closure (GCI) and glottal openings (GOI) are accurately estimated from LP residual [5]. Epoch extraction by computing the variations of amplitude envelopes of the frequency shifted speech, filtered through a single pole filter, is proposed in the SFF method. The SFF method is proposed mainly to avoid the artifacts involved in the block processing stages of the former methods [7], [8].

Apart from aforesaid conventional methods, there exist a few methods for epoch extraction by decomposing the given speech into some of the meaningful component modes. Rajib et al. proposed effectiveness of intrinsic mode functions (IMFs), obtained by empirical mode decomposition (EMD) of speech, in carrying epochal information [9]. Variational mode decomposition is used to decompose the given signal into components modes with better frequency resolution than EMD [10]. The component mode whose center frequency is close to the average fundamental frequency is reported to provide accurate estimation of epochs in speech compared to the conventional epoch extraction methods [11]. All EMD and VMD based time domain decomposition based techniques ensures reduce deviation of the estimated epochs with respect to the reference epochs by the accurate time frequency representation.

Figure 1 compares the time frequency plot obtained for a quadratic chirp signal, with a starting frequency of 500 Hz (at time $t=0$) reducing to 100 Hz at $t=2$ and again increasing to 500 Hz at $t=4$, with short time fourier transform (STFT), continuous wavelet transform (CWT) and wavelet synchrosqueezed transform (WSST). Due to fixed analysis frame size, the time-frequency representation obtained from STFT has poor time and frequency resolution (seen as the thick smoothed time frequency plot). Orthogonal basis function obtained by the shifting and scaling of the chosen mother wavelets give better time and frequency resolution as compared to the STFT [12]. The time evolutions of the frequencies are clearly visible as step like discontinuities in Figure 1 (b). Compared to Figure 1 (a) & (b) the subplot (c) obtained from WSST shows time-frequency representation with improved time and frequency resolutions. The tight time and frequency evolutions can be clearly seen in the WSST of the chirp signal as compared to STFT and CWT [13], [14]. Accurate time-frequency representation obtained in the wavelet synchrosqueezed transform is exploited for the epoch estimation work presented in this paper. Accordingly the work presented in this paper is organized in

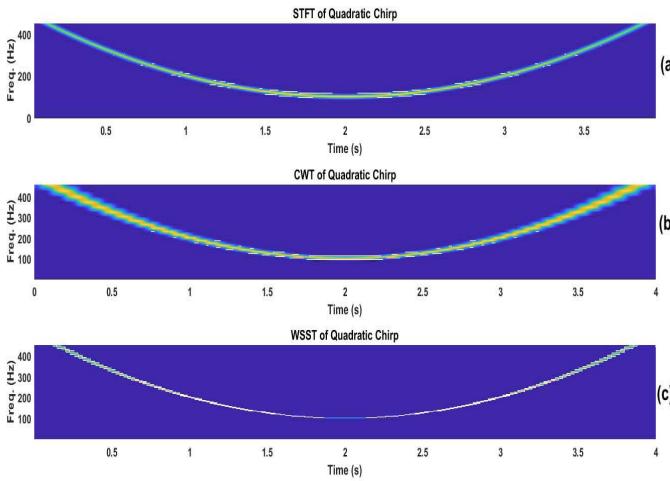


Fig. 1: Comparing the Time-Frequency resolution of a quadratic chirp signal using (a) STFT, (b) continuous wavelet transforms(CWT) and (c) wavelet synchrosqueezed transforms(WSST). (subplots (a) and (b) are regenerated from the quadratic chirp time-frequency example given in the Matlab wavelet toolbox)

the following way. Section II provides the theoretical details of wavelet synchrosqueezed transforms. The proposed method of epoch estimation is explained in Section III. Experimental studies and results are discussed in Section IV and finally sectionsummary summarizes the work.

II. WAVELET SYNCHROSQUEEZED TRANSFORMS FOR IMPROVED TIME AND FREQUENCY REPRESENTATION

The time frequency representation by continuous wavelet transform is given by the Equation 1.

$$W_x(a, b) = \int \frac{1}{\sqrt{a}} x(t) \psi\left(\frac{t-b}{a}\right) dt \quad (1)$$

where a and b are the respective scaling and shifting variables and $\psi(t)$ is the chosen mother wavelet. The frequency resolution and time resolution of CWT representation is determined by a and b scales in the continuous domain, respectively. The instantaneous frequency, $\omega_x(a, b)$, computed from CWT is given in Equation 2

$$\omega_x(a, b) = -i \frac{1}{W_x(a, b)} \frac{\partial}{\partial b} W_x(a, b) \quad (2)$$

For the analysis of auditory signals, Daubechies et al. observed that the frequency resolution of the CWT is centered around the center frequency ω_0 of chosen wavelet in the positive frequency axis. Hence, the frequency evolutions in $W_x(a, b)$ show a spread around the center line, $a=\omega_0/\omega$, of the time scale axis. This spread is observed in the analysis of the mono component signal, $x(t)=A\cos(\omega t)$ in [13]. As proposed in [13], to avoid the frequency spread in the time-frequency analysis of mono components, the scale variables (b, a) of the CWT ($W_x(a, b)$) is mapped to the new time-frequency plane represented by $(b, \omega_s(a, b))$ where $\omega_s(a, b)$ is the instantaneous frequency estimated from $W_x(a, b)$. This is termed

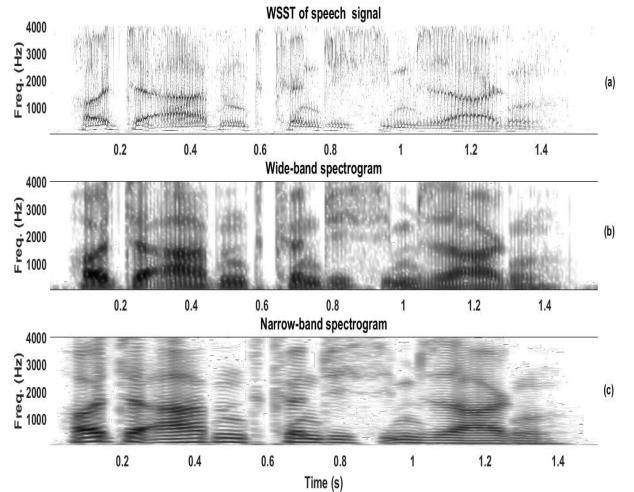


Fig. 2: Comparing the Time-Frequency representation of a speech signal (a) WSST, (b) wide-band spectrogram obtained from STFT and (c) narrow-band spectrogram obtained from STFT

as synchrosqueezing. In wavelet synchrosqueezed transform, $W_x(a, b)$ is computed for discrete values of scale variable a_k s forming bins of scale variable a and frequency variable ω . The synchrosqueezed transform $T(\omega, b)$ is computed at ω_l by centering around the successive frequency bins $[\omega_l - \frac{1}{2}\Delta\omega, \omega_l + \frac{1}{2}\Delta\omega]$ and is given by,

$$T_x(\omega_l, b) = \frac{1}{\Delta\omega} \sum_{a_k: |\omega(a_k, b) - \omega_l| \leq \frac{\Delta\omega}{2}} W_x(a_k, b) a_k^{-3/2} (\Delta a)_k \quad (3)$$

where $(\Delta a)_k$ is the difference between discrete scales of successive bins and is represented as $(\Delta a)_k = a_k - a_{k-1}$. The reallocation of original time-frequency scale to a different time-frequency plane indeed helps to sharpen the time-frequency representation of the given signal.

After synchrosqueezing each of the time scale can be reconstructed using the following synthesis formulae,

$$x(b) = \operatorname{Re} \left[C_\psi^{-1} \int_0^\infty W_x(a, b) a^{-3/2} da \right] \quad (4)$$

where $C_\psi = \int_0^\infty \overline{\psi(\xi)} \frac{d\xi}{\xi}$ and the chosen mother wavelet ψ has positive frequency axis ($\psi(\xi) \geq 0$).

III. EVIDENCE OF EPOCHAL INFORMATION IN WAVELET SYNCHROSQUEEZED TRANSFORM OF SPEECH

In the proposed method, the wavelet synchrosqueezing transform is used as an EMD or VMD like tool to decompose the given speech signal into a number of component signals corresponding each nonlinear frequency scale. Figure 2 plots the time-frequency magnitude plot of speech signal obtained using WSST and spectrograms (narrow-band and wide-band)

TABLE I: Performance evaluation of WSST based epoch extraction from speech on CMU Arctic database.

Method	IDR (%)	MR (%)	FAR (%)	IDA (msec)
ZFF	99.34	0.04	0.62	0.35
WSST	93.46	0.08	6.45	0.22
ZFF-WSST	99.34	0.04	0.62	0.31

computed using STFT. Time-Frequency representation obtained using WSST as plotted in Figure 2(a) shows sharp tracking of the formants, pitch and higher harmonics in the speech signal. The wide-band and narrow-band representation show relatively wider spread of the formants (from wide band spectrogram), pitch and higher harmonic (from narrow-band) tracks

As the context of the present work is epoch estimation, the instantaneous F_0 information present in the lower frequency component modes have to be explored. To extract the instantaneous F_0 , the time series correspond to the higher and extremely lower frequencies (typically frequency component modes in the human pitch range) are ignored for the reconstruction of the signal from WSST transform. The reconstructed signal is observed to be free from higher frequency vocal tract interactions and higher harmonics. The epochs are hypothesized as the positive zero crossings in reconstructed signal obtained from the inverse transform of modified WSST. The proposed algorithm for epoch estimation from speech signals is carried out in the following four steps.

- Computing the WSST transform of speech signal
- The component modes correspond to those frequencies that are within the human pitch range around the average pitch frequency are retained to form the modified WSST
- Signal reconstructed by taking the inverse of the modified WSST
- Zero crossings computed from the reconstructed signal are hypothesized as the epochs in speech

Figure 3 indicates the evidence of epochal information in the signal reconstructed by taking the inverse of the modified WSST of the given speech signal. The positive zero crossings that are indicated as red colored arrows in Figure 3(b) fall aligned with the dominant discontinuities in the differenced EGG (in Figure 3(c)) which are referred as the ground truth epochal information.

IV. EXPERIMENTAL RESULTS

A. Performance Evaluation of WSST based Epoch Estimation

The epoch estimation performance of the proposed method is evaluated for the studio quality clean speech signals on the phonetically balanced CMU-Arctic database with simultaneous speech and EGG recordings [15]. The epoch estimation performance is evaluated based on measures such as epoch identification rate (IDR), miss rate (MR), false alarm rate (FAR) and epoch identification accuracy (IDA). The details of these measures are provided in [1], [5], [6], [16]. The Table I shows epoch IDR, MR, FAR and IDA measures obtained from the proposed WSST based epoch estimation method from

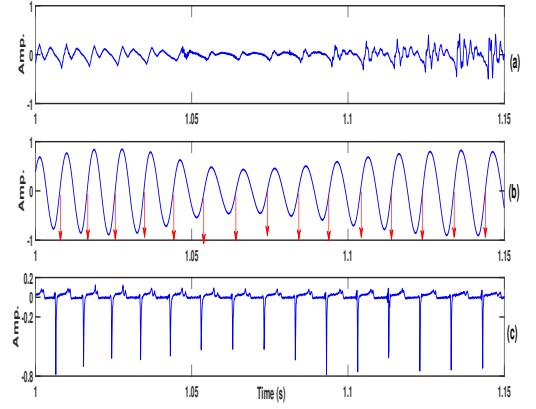


Fig. 3: Evidence of epochs in the reconstructed signal from modified WSST. (a) Voiced segment of speech, (b) Reconstructed signal from WSST and (c) the differenced EGG signal showing the reference epochs as dominant discontinuities. The estimated epochs are indicated as red arrows at the positive zero crossings in the reconstructed signal.

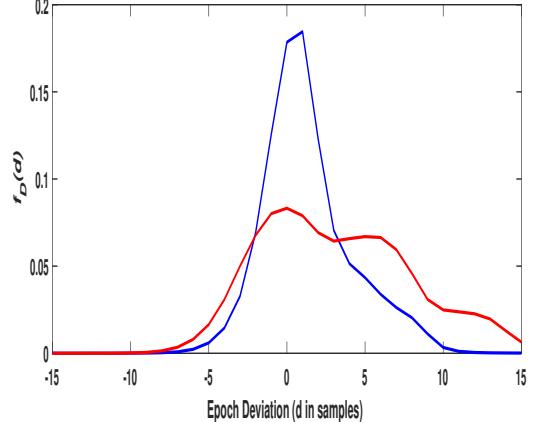


Fig. 4: The probability density functions ($f_D(d)$) of the epoch deviations for the epochs estimated using proposed WSST method and the conventional ZFF method. The blue color plot indicates WSST method and 'red' one indicates the ZFF method

speech. The average epoch performance measures obtained from each speaker are comparable with that of the popular methods like ZFF [4], DYPSA [1] and IPR [6] epoch estimation methods. For comparative performance analysis, the epoch estimation performance of ZFF method is also provided in the Table I. Based on the performance comparison, proposed method showed slightly degraded performance in terms of IDR, MR and FAR. However, the epoch identification accuracy obtained for the proposed method is better than that of the ZFF method. To measure the significance of the epoch identification accuracy, the probability distributions of the deviations obtained from the estimated epochs are plotted for both the methods in Figure 4. The events represented by the random variable D correspond to the actual deviations in

TABLE II: *Effectiveness of WSST in estimating epochs from Emotive Speech Signals in EmoDb. The acronym ZFF-WSST corresponds to the method of applying WSST on ZFFS.*

Emotion	IDR (%)			MR (%)			FAR (%)			IDA (ms)		
	WSST	WSST – ZFF	ZFF	WSST	WSST – ZFF	ZFF	WSST	WSST – ZFF	ZFF	WSST	WSST – ZFF	ZFF
Anger	93.46	96.48	91.73	1.90	3.04	1.90	5.45	0.48	6.37	0.35	0.32	0.35
Happy	94.62	96.12	96.15	1.32	3.55	1.89	4.06	0.33	3.55	0.35	0.32	0.33
Fear	98.02	97.80	97.60	0.00	2.12	1.62	0.01	0.08	0.79	0.28	0.26	0.27
Boredom	98.57	98.44	99.36	0.47	1.36	0.44	0.96	0.20	0.20	0.26	0.33	0.34
Neutral	98.36	99.19	99.45	0.63	0.43	0.71	1.01	0.11	0.09	0.23	0.28	0.29
Average	96.60	97.60	96.85	0.84	2.10	1.31	2.29	0.24	0.20	0.29	0.30	0.32

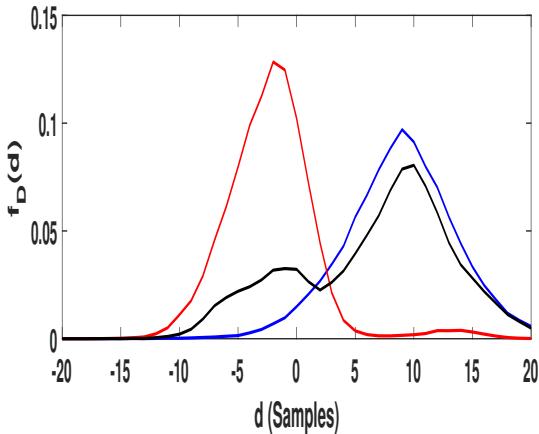


Fig. 5: The probability density functions ($f_D(d)$) of the epoch deviations for the epochs estimated by WSST of Speech (red colored plot), conventional ZFF (blue colored plot) and WSST of ZFFS (Black colored plot) for German Emotion speech database (EmoDb).

samples with respect to the corresponding reference epochs obtained from the EGG. From the Figure 4, the probabilities of the estimated epoch with reduced deviations are higher (reduced variance of p.d.f) for WSST method as compared to conventional ZFF method. Also, Table I shows the effectiveness of WSST in improving the epoch estimation performance of the conventional ZFF method. Instead of speech signal, the zero frequency filtered signal (ZFFS) is subjected to WSST transform. The reduced epoch deviation indicates improved epoch identification accuracy of the ZFF-WSST approach. As the WSST sharpens the instantaneous frequency variations present in the ZFFS, the epochs are estimated with better resolution as shown in the Table I. The WSST of ZFFS is expected to provide better estimation performance for emotion speech signals where the instantaneous pitch is varying rapidly.

B. Effectiveness of WSST in estimating epochs from Emotive Speech Signals

Due to rapid pitch variations in the emotive speech signals, reliable epoch estimation from emotive utterances is always challenging [8], [17]. In conventional ZFF method, the local mean subtraction using fixed window length fails to capture rapid instantaneous pitch variations present in emotive utterances [8]. WSST is then applied to ZFFS obtained from the conventional ZFF method to sharpen the instantaneous pitch variations present in the emotive utterances. The smooth ZFFS with less spurious zero crossings improves epoch estimation

performance. German emotional speech database (EmoDb) having simultaneous EGG recordings for each utterances is used to verify the effectiveness of the proposed WSST based epoch estimation from emotive speech signals. Five emotions simulated by 10 professional speakers in 10 texts with average 100 utterances per emotion of EmoDb are used for the performance evaluation. Originally 16 kHz sampled signals available in EmoDb are resampled to 8 kHz for epoch estimation. From the Table II, the epoch estimation performance obtained for ZFF-WSST is much better than that of ZFF method in terms of IDR, MR, FAR and IDA. Many post processing methods are proposed for refining the zero frequency filtered signal to improve epoch estimation from emotive speech signals [17]–[20]. The WSST of speech is also showed better epoch estimation performance as compared to conventional ZFF in an average sense. The WSST based method is free from setting additional block processing related parameters such as window length as in the case of conventional ZFF method. The results obtained in Table II indicate that the proposed method showed better epoch estimation performance for emotive speech signals in the absence of any postprocessing technique which is essential for conventional ZFF approach. Figure 5 plots the probability density functions of the epoch deviations in WSST of speech (red colored plot), conventional ZFF (blue colored plot) and WSST of ZFFS plot (black colored plot) obtained for all the five emotion utterances from EmoDb. Compared to clean speech signals, more spread is observed for all the distributions obtained for emotions. for the emotion case also, the epochs obtained from WSST of speech signals provided most number of minimum epoch deviations. Even though the spread for the density functions of deviations obtained by taking the WSST of ZFFS showed more spread than that of the conventional ZFF, an increased probability density of epochs (around ± 3 samples) are observed at lower epoch deviations.

V. SUMMARY & CONCLUSIONS

The present work proposed wavelet synchrosqueezed transform (WSST) for improving the epoch estimation accuracies from speech signals. The time-frequency sharpening properties of the WSST are exploited for improving the epoch estimation accuracies from speech. Even though, the epoch estimation using WSST of speech showed degradation in epoch identification rate, miss and false alarm rates, the identification accuracy is found to be significantly better than the conventional methods. The tightening of time-frequency representation of the instantaneous frequency present in the signal is motivated us to subject zero frequency filtered signal obtained from the conventional ZFF method to further improve the epoch estimation performance. The WSST of ZFFS (ZFF-WSST) found to provide significantly better performance for clean

speech signals as well as emotive speech. The future work should focus on the effective use of WSST method in other degraded conditions such as telephonic quality, in the presence of noise and etc. where conventional methods show significant degradation in estimating epochs.

REFERENCES

- [1] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using DYPSA algorithm," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007.
- [2] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 4, pp. 325–333, Sep.1995.
- [3] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using hilbert envelope and group delay function," *IEEE Signal Processing Letters*, vol. 14, pp. 762–765, Oct. 2007.
- [4] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1614, Nov. 2008.
- [5] T. Drugman and T. Dutoit, "Glottal closure and opening instant detection from speech signals," in *Proc. INTERSPEECH*, 2009, pp. 2891–2895.
- [6] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio Speech and Language Process.*, vol. 21, no. 12, pp. 2471–2480, 2013.
- [7] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 23, no. 4, pp. 705–717, Apr. 2015.
- [8] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.
- [9] S. Rajib, S. R. M. Prasanna, H. L. Rufiner, and G. Schlotthauer, "Detection of the glottal closure instants using empirical mode decomposition," *International Journal of Circuits, Systems and Signal Processing*, pp. 1–29, 2017.
- [10] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. on Signal Processing*, vol. 62, no. 3, pp. 531–544, Feb. 2014.
- [11] G. J. Lal, E. A. Gopalakrishnan, and D. Govind, "Accurate estimation of glottal closure instants and glottal opening instants from electroglottographic signal using variational mode decomposition," *International Journal of Circuits, Systems and Signal Processing*, vol. 37, no. 2, pp. 810–830, Feb. 2018.
- [12] L. Cohen, *Time-Frequency Analysis: Theory and Applications*, S. P. Series, Ed. ser. Signal Processing Series. Englewood Cliffs: Prentice-Hall, 1995.
- [13] I. Daubechies and S. Maes, *A nonlinear Squeezing of the continuous wavelet transforms based on auditory nerve model*, ser. Wavelets in medicinal biology, M. U. E. A. Aldroubi, Ed. CRC Press, 1996.
- [14] I. Daubechies, J. Lu, and H.-T. Wu, "Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool," *Appl. Comput. Harmon. Anal.*, vol. 30, pp. 243–261, 2011.
- [15] J. Kominek and A. Black, "CMU-Arctic speech databases," in *in 5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224.
- [16] D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *proc. INTERSPEECH 2011*, Aug. 2011.
- [17] D. Govind and S. R. M. Prasanna, "Epoch extraction from emotional speech," in *in proc. Signal Procesing & Communications (SPCOM)*, July 2012, pp. 1–5.
- [18] D. Pravena and D. Govind, "Expressive speech analysis for epoch extraction using zero frequency filtering approach," in *Proc. IEEE International Student Symposium*. Indian Institute of Technology Kharagpur, 2016.
- [19] S. R. Kadiri and B. Yegnanarayana, "Analysis of singing voice for epoch extraction using zero frequency filtering method," in *International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [20] D. Govind, D. Pravena, and G. Ajay, "Improved epoch extraction using variational mode decomposition based spectral smoothing of zero frequency filtered emotive speech signals," in *Proceding of National Conference on Communications (NCC)*. Indian Institute of Technology Hyderabad, Feb 2018.

Performance Analysis and Optimization of Interference limited Multi-Antenna BRN

Imtiyaz Khan*, Krishna Kanth Dhulipudi † and Poonam Singh‡

Department of Electronics and Communication Engineering, NIT Rourkela, India

Email: *imtiyazfaith@gmail.com, †krishnakanthdulipudi@gmail.com, ‡psingh@nitrkl.ac.in

Abstract—This paper investigates the outage performance of multiple antenna bidirectional relaying network (BRN) in the presence of co-channel interference (CCI). Herein, multi-antenna sources exchange information bi-directionally with the help of a single-antenna relay terminal. Under such scenario, we evaluate and compare the performance of two amplify-and-forward based multi-antenna transmission strategies viz., beamforming (BF) and antenna selection (AS). We derive the tight upper bound expressions of end outage probability (OP) for both the strategies over Rayleigh fading channel. We further conduct asymptotic analysis to examine the achievable diversity order of the considered system. To gain more insights, we analyze the power optimization problem to minimize the OP for different scenarios. Finally, Monte-Carlo simulation results are given to attest our theoretical analysis. Our finding suggests that the BF overperform the AS scheme at the expense of additional complexity.

Index Terms—Bidirectional relay systems, co-channel interferences, multi-antenna, AF relaying, outage probability, asymptotic analysis.

I. INTRODUCTION

Reliability, reduced power consumption and wider coverage are the primary concerns of future wireless systems. Relaying communication is the promising candidate to fulfill these requirements. Among the relaying schemes, unidirectional relaying network (URN) is quite popular because of its simplicity and low implementation complexity. On the other hand, bidirectional relaying network (BRN) has recently gained significant popularity due to their higher spectral efficiency as compared to URN [1]. To further enhance the communication reliability, multiple-input-multiple-output (MIMO) technologies can be used effectively. Specifically, beamforming (BF) and Tx/Rx antenna selection (AS) are used as a transmission strategy in multi-antenna systems. Maximal ratio transmission (MRT) and maximal ratio combining (MRC) are included in the BF technique. Further, AS eliminates the need of multiple transmit and receive radio frequency (RF) chains. Therefore AS scheme reduces the complexity, power requirements and the cost of the MIMO transceivers [2].

However, in the cellular communication scenario, co-channel interference (CCI) is the dominant limiting factor due to its frequency reuse strategy. The performance of the relaying system is also degraded in the presence of CCI and limited diversity gains are achieved [3]. In MIMO relaying systems, each RF chain suffers from interferences and effect becomes manifolds. Therefore, the study of CCI in multiantenna relay

network has its prime importance so that, this paper presents the comparative analysis of such system.

Prior related research: In [4], approximate analysis of interference limited BRN is evaluated in a Rayleigh fading environment. It is further investigated in [3], where authors studied the bounds on system performance in Nakagami-m fading. Recently, the effect of CCI is evaluated in decode-and-forward (DF) based spectrum sharing BRN system in Nakagami-m fading [5]. Authors in [6] investigate the joint impact of hardware impairment and CCI in the system performance of DF-URN, where N^{th} best relay selection employed and CCI is assumed only at the relay.

Only a few studies, [7]–[11], addressing the issue of CCI with the multi-antenna system. In [7] authors addressed performance comparison of different beamforming technique in the presence of CCI in URN but interference is assumed only at the multi-antenna relay node. Li *et al.* in [8], evaluates the multiantenna system performance of BF based URN where interference is assumed only at the relay terminal. Recently, authors in [9] proposed optimal beamforming in the multi-antenna system in the presence of CCI but the analysis is limited to URN. The issue of CCI in multi-antenna BRN is studied in [10], where beamforming is used among the terminals and the individual outage probability (OP) is evaluated in a Rayleigh fading environment. Khan *et al.* in [11] have analyzed the overall performance of antenna selection (AS) based BRN in the presence of a finite number of CCIs only at the relay node.

Motivation and Contribution: Firstly, [2] considered multiantenna system but do not consider the CCIs. Secondly, [10] and [11] have taken into account both CCIs and multi-antenna; while interference assumed only at the relay. However, [3], [4] and this paper considered into account the CCIs at both the source nodes and the relay node, which is more general scenario than other existing works; thirdly, [3] and [4] don't employ the multi-antenna system. Another key limitation of above-mentioned literature is that, equal power interference is assumed, which is impractical because different devices with different transmitting power can interfere with the intended device. Therefore, to the best of our knowledge, analysis of multi-antenna BRNs with CCI of unequal powers at all the terminals have not yet been studied. Thus, this paper fills this gap by deriving the comparative performance analysis of two multi-antenna transmission schemes viz., BF and AS in the presence of CCI at all the terminals. We have also investigated the power optimization technique to minimize OP for different

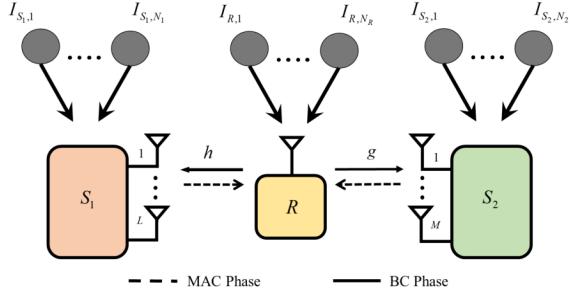


Fig. 1. Multi-antenna two-way AF relaying

interference power profile.

Notations: In this paper scalars and vectors are represented as italic symbols and lower case boldface symbols respectively. For any scalar a , absolute value is denoted by $|a|$. For a given complex vector \mathbf{a} , $(\mathbf{a})^T$ represents the transpose, $(\mathbf{a})^\dagger$ represent the conjugate transpose and $\|\mathbf{a}\|$ denotes the Euclidean norm. $\mathcal{CN}(\mu, \sigma^2)$ denotes a complex circular Gaussian random variable with mean μ and variance σ^2 . where $\Gamma(\cdot)$ denote the gamma function [12, Eq. (8.310.1)]. $\mathbb{E}[\cdot]$ shows the expectation.

II. SYSTEM MODEL AND TRANSMISSION SCHEME

We consider a two-way relaying system where source S_1 and S_2 exchange information via intermediate relay R as shown in Fig. 1. The direct link between source S_1 and S_2 is absent due to heavy shadowing or large distance between terminals. We assume that S_1 and S_2 are equipped with L and M antennas respectively, while R is equipped with a single antenna¹. Further, we assume that all the terminals are operating in half-duplex mode. The communication between nodes take place in two phases namely multiple access (MAC) phase and broadcast (BC) phase respectively. In MAC phase S_1 and S_2 communicate with R , afterwards R amplifies the received signal and forwards to S_1 and S_2 in BC phase. Furthermore, we assume that all the terminals are impaired by inter and intra channel interferences from other users in the network. From Fig. 1, S_1 , S_2 and R are inflicted by N_1 , N_2 and N_R interferers respectively. We consider that the channels are reciprocal and channel amplitude undergo flat Rayleigh fading. Now, the system modeling for respective transmission schemes are as follows

A. Beamforming

Source S_1 transmits u_1 to R using L antennas and S_2 transmits u_2 to R using M antennas. In MAC phase, signal received at relay terminal is given as

$$y_R = \sqrt{P_1} \mathbf{h}^T \mathbf{w}_1 u_1 + \sqrt{P_2} \mathbf{g}^T \mathbf{w}_2 u_2 + \sum_{i=1}^{N_R} \sqrt{P_{R,i}} f_{R,i} u_{R,i} + n_R, \quad (1)$$

¹The considered network has various practical applications, e.g., D2D communications where two devices with multi-antenna exchange information with the help of relay node and wireless sensor networks where a single-antenna mobile relay help two multiple-antenna access points to exchange information.

where u_i and P_i denote the respective transmit signal (with unit energy) and transmit power from the source S_i , for $i \in \{1, 2\}$, and n_R represents the additive white Gaussian noise (AWGN) at R . $\mathbf{h} = [h_1, h_2, \dots, h_L]^T$ and $\mathbf{g} = [g_1, g_2, \dots, g_M]^T$ defines the channel fading coefficient between $S_1 \rightarrow R$ and $S_2 \rightarrow R$ respectively. $\mathbf{w}_1 = (\mathbf{h}^\dagger / \|\mathbf{h}\|)^T$ and $\mathbf{w}_2 = (\mathbf{g}^\dagger / \|\mathbf{g}\|)^T$ defines the transmit weight vectors at S_1 and S_2 respectively. The transmit power from S_1 and S_2 are given as $\mathbb{E}[\|\sqrt{P_1} \mathbf{w}_1\|^2] = P_1$ and $\mathbb{E}[\|\sqrt{P_2} \mathbf{w}_2\|^2] = P_2$, respectively. Besides, $\{f_{R,i}\}_{i=1}^{N_R}$ represent channel coefficient between R and i -th interferer, $P_{R,i}$ is the received power of i -th interferer at R , $u_{R,i}$ is the signal transmitted by i -th interferer. In BC phase, the received signal at S_1 and S_2 are given as

$$y_1 = \mathbf{w}_1^\dagger \left(\mathcal{G}_{BF} \mathbf{h} y_R + \sum_{j=1}^{N_1} \sqrt{P_{1,j}} \mathbf{f}_{1,j} u_{1,j} + n_1 \right), \quad (2)$$

$$y_2 = \mathbf{w}_2^\dagger \left(\mathcal{G}_{BF} \mathbf{g} y_R + \sum_{k=1}^{N_2} \sqrt{P_{2,k}} \mathbf{f}_{2,k} u_{2,k} + n_2 \right), \quad (3)$$

respectively, where, $\mathbf{f}_{1,j} = [f_{1,1}, f_{1,2}, \dots, f_{1,N_1}]$ and $\mathbf{f}_{2,k} = [f_{2,1}, f_{2,2}, \dots, f_{2,N_2}]$ represent the channel fading coefficient between S_1 & j -th interferer and S_2 & k -th interferer respectively, n_1 and n_2 represents AWGN at S_1 and S_2 respectively and \mathcal{G}_{BF} defines the variable gain of relay which is given as² $\mathcal{G}_{BF}^2 = P_R / (P_1 \|\mathbf{h}\|^2 + P_2 \|\mathbf{g}\|^2 + \sum_{i=1}^{N_R} P_{R,i} |f_{R,i}|^2 + N_0)$.

Now, substituting (1) in (2) and after canceling the self interference term we get

$$\tilde{y}_1 = \mathcal{G}_{BF} \sqrt{P_2} \mathbf{w}_1^\dagger \mathbf{h} \mathbf{g}^\dagger \mathbf{w}_2 u_2 + \mathbf{w}_1^\dagger \sum_{j=1}^{N_1} \sqrt{P_{1,j}} \mathbf{f}_{1,j} u_{1,j} + \mathcal{G}_{BF} \mathbf{w}_1^\dagger \mathbf{h} n_R + \mathbf{w}_1^\dagger \mathbf{n}_1. \quad (4)$$

Since the entries of $\mathbf{f}_{1,j}$ are independent and identically distributed (i.i.d) complex Gaussian RVs, the fading coefficient of each interferer is still a complex Gaussian RV even after the MRC [13]. Thus to facilitate further analysis, we can express \tilde{y}_1 equally as follows

$$\tilde{y}_1 = \mathcal{G}_{BF} \sqrt{P_2} \mathbf{w}_1^\dagger \mathbf{h} \mathbf{g}^\dagger \mathbf{w}_2 u_2 + \mathbf{w}_1^\dagger \sum_{j=1}^{N_1} \sqrt{P_{1,j}} f_{1,j} u_{1,j} + \mathcal{G}_{BF} \mathbf{w}_1^\dagger \mathbf{h} n_R + \mathbf{w}_1^\dagger \mathbf{n}_1, \quad (5)$$

similarly we will get \tilde{y}_2 as

$$\tilde{y}_2 = \mathcal{G}_{BF} \sqrt{P_1} \mathbf{w}_2^\dagger \mathbf{h}^\dagger \mathbf{g} \mathbf{w}_1 u_1 + \mathbf{w}_2^\dagger \sum_{j=1}^{N_2} \sqrt{P_{2,j}} f_{2,j} u_{2,j} + \mathcal{G}_{BF} \mathbf{w}_2^\dagger \mathbf{g} n_R + \mathbf{w}_2^\dagger \mathbf{n}_2. \quad (6)$$

Thus the end-to-end signal-to-interference-noise-ratio (SINR) from S_ρ to S_ω in generalized form is given by

$$\gamma_{S_\rho \rightarrow S_\omega}^{BF} = \frac{\xi \gamma_\rho \gamma_\omega}{\xi \gamma_\omega \gamma_R + (\gamma_\rho + \gamma_\omega) \gamma_\lambda + \gamma_R \gamma_\lambda}, \quad (7)$$

we have assumed equal power³ $P_1 = P_2 = P$ then $\xi = \frac{P_R}{P}$. Here $(\rho, \omega, \lambda) \in \{(1, 2, \eta), (2, 1, \delta)\}$, $\gamma_\eta = U + 1$, $\gamma_R = V + 1$ and $\gamma_\delta = W + 1$, where $U = \sum_{k=1}^{N_2} \frac{P_{2,k} |f_{2,k}|^2}{N_0}$,

²Without loss of generality we have assumed that the noise at all the terminal (n_R , n_1 , n_2) follows $\mathcal{CN}(0, N_0)$.

³The assumption of equal power at SU terminals doesn't make any loss of generality in system performance analysis as the differing transmit power can be easily included in dissimilar average fading power.

$$V = \sum_{i=1}^{N_R} \frac{P_{R,i}|f_{R,i}|^2}{N_0} \text{ and } W = \sum_{j=1}^{N_1} \frac{P_{1,j}|f_{1,j}|^2}{N_0}. \text{ Here } \gamma_1 = \bar{\gamma} \|\mathbf{h}\|^2, \gamma_2 = \bar{\gamma} \|\mathbf{g}\|^2 \text{ and } \bar{\gamma} = \frac{P}{N_0}.$$

B. Antenna Selection

From multiple antenna, one is selected which maximizes the instantaneous SINR at R [2]. Thus in MAC phase, the received signal at R is

$$y_R = \sqrt{P_1} h_{l^*} u_1 + \sqrt{P_2} g_{m^*} u_2 + \sum_{i=1}^{N_R} \sqrt{P_{R,i}} f_{R,i} u_{R,i} + n_R, \quad (8)$$

where $|h_{l^*}| = \max_{1 \leq l \leq L} |h_l|$ and $|g_{m^*}| = \max_{1 \leq m \leq M} |g_m|$ denotes the magnitude of the fading coefficient between S₁ & R, and the magnitude of the fading coefficient between S₂ & R of selected antenna respectively.

In the BC phase, R applies the scaling gain \mathcal{G}_{AS} to y_R and forwards it to both S₁ and S₂ with power P_R . Thus the received signal at S₁ and S₂ terminal are given as

$$y_1 = \mathcal{G}_{AS} h_{l^*} y_R + \sum_{j=1}^{N_1} \sqrt{P_{1,j}} f_{1,j} u_{1,j} + n_1, \quad (9)$$

$$y_2 = \mathcal{G}_{AS} g_{m^*} y_R + \sum_{j=1}^{N_2} \sqrt{P_{2,j}} f_{2,j} u_{2,j} + n_2, \quad (10)$$

where, G defines the variable gain of relay which is given as $\mathcal{G}_{AS}^2 = P_R / (P_1 |h_{l^*}|^2 + P_2 |g_{m^*}|^2 + \sum_{i=1}^{N_R} P_{R,i} |f_{R,i}|^2 + N_0)$. After canceling the self interference term from y_1 and y_2 leads to

$$\begin{aligned} y_1^* &= \mathcal{G}_{AS} \sqrt{P_2} h_{l^*} g_{m^*} s_2 + \mathcal{G}_{AS} h_{l^*} \sum_{i=1}^{N_R} \sqrt{P_{R,i}} f_{R,i} u_{R,i} \\ &\quad + \sum_{j=1}^{N_1} \sqrt{P_{1,j}} f_{1,j} u_{1,j} + \mathcal{G}_{AS} h_{l^*} n_R + n_1, \end{aligned} \quad (11)$$

$$\begin{aligned} y_2^* &= \mathcal{G}_{AS} \sqrt{P_1} h_{l^*} g_{m^*} s_1 + \sum_{j=1}^{N_2} \sqrt{P_{2,j}} f_{2,j} u_{2,j} \\ &\quad + \mathcal{G}_{AS} g_{m^*} \sum_{i=1}^{N_R} \sqrt{P_{R,i}} f_{R,i} u_{R,i} + \mathcal{G}_{AS} g_{m^*} n_R + n_2. \end{aligned} \quad (12)$$

By following the same assumption in BF, the end-to-end SINR from S_p to S_ω in generalized form is given by

$$\gamma_{S_p \rightarrow S_\omega}^{AS} = \frac{\xi \tilde{\gamma}_p \tilde{\gamma}_\omega}{\xi \tilde{\gamma}_\omega \tilde{\gamma}_R + (\tilde{\gamma}_p + \tilde{\gamma}_\omega) \gamma_\lambda + \tilde{\gamma}_R \gamma_\lambda}, \quad (13)$$

where $\tilde{\gamma}_1 = \bar{\gamma} |h_{l^*}|^2$, $\tilde{\gamma}_2 = \bar{\gamma} |g_{m^*}|^2$. Thus, based on (7) and (13), the OP is formulated in next section for beamforming and antenna selection respectively.

III. OUTAGE PROBABILITY

In this section, we analyze the performance of the proposed system model. The OP expressions for respective transmission schemes are derived as follows:

A. Beamforming

Proposition 1: From (7), the cdf of instantaneous SINR for BF scheme is given as

$$\begin{aligned} \mathcal{P}_{out}^{BF} &= 1 - e^{-\gamma_{th} \mathcal{A}} \sum_{i,j,k,l,n} \sum_{a,b,c,d} \frac{\xi^{k-i-j} \binom{j}{k} \binom{k}{n}}{\Omega_h^i \Omega_g^j i! j!} \left(\frac{\gamma}{\bar{\gamma}} \right)^{i+j} \\ &\quad \times \frac{\mathcal{X}_{a,b}(\mathcal{Q}_1) \mathcal{X}_{c,d}(\mathcal{Q}_2) \Gamma(l+b) \Gamma(n+d)}{(\mu_{[a]})^b (\kappa_{[c]})^d \Gamma(b) \Gamma(d)} \\ &\quad \left(\frac{1}{\mu_{[a]}} + \gamma_{th} \mathcal{B} \right)^{-l-b} \left(\frac{1}{\kappa_{[c]}} + \frac{\gamma_{th}}{\bar{\gamma} \Omega_g} \right)^{-n-d}, \end{aligned} \quad (14)$$

where $\mathcal{A} = (\Omega_g + \Omega_h(\xi + 1)) / (\xi \bar{\gamma} \Omega_h \Omega_g)$, $\mathcal{B} = (\Omega_g + \Omega_h) / (\xi \bar{\gamma} \Omega_h \Omega_g)$, $\Omega_h = \mathbb{E}\{||\mathbf{h}||^2\}$, $\Omega_g = \mathbb{E}\{||\mathbf{g}||^2\}$ with $\sum_{i,j,k,l,n} \sum_{a,b,c,d}$ being the short-hand notation for $\sum_{i=0}^{L-1} \sum_{j=0}^{M-1} \sum_{k=0}^j \sum_{l=0}^{i+j-k} \sum_{n=0}^k$.

Proof: See Appendix.

B. Antenna Selection

Proposition 2: From (13), the cdf of instantaneous SINR for AS scheme is given as

$$\mathcal{P}_{out}^{AS} = \mathcal{F}_1 + \mathcal{F}_2 - \mathcal{F}_3, \quad (15)$$

where

$$\mathcal{F}_1 = \sum_{i=0}^L \sum_{a,b} \binom{L}{i} \frac{\mathcal{X}_{a,b}(\mathcal{Q}_1) (-1)^i e^{-\frac{\gamma_{th} i}{\bar{\gamma} \xi \Omega_h}}}{(\mu_{[a]})^b} \left(\frac{i \gamma_{th}}{\bar{\gamma} \xi \Omega_h} + \frac{1}{\mu_{[a]}} \right)^{-b}, \quad (16)$$

$$\begin{aligned} \mathcal{F}_2 &= \sum_{j=0}^M \sum_{a,b,c,d} \binom{M}{j} (-1)^j e^{-\frac{\gamma_{th} j (\xi+1)}{\bar{\gamma} \xi \Omega_g}} \\ &\quad \frac{\mathcal{X}_{a,b}(\mathcal{Q}_1) \mathcal{X}_{c,d}(\mathcal{Q}_2)}{(\mu_{[a]})^b (\kappa_{[c]})^d} \left(\frac{j \gamma_{th}}{\bar{\gamma} \xi \Omega_g} + \frac{1}{\mu_{[a]}} \right)^{-b} \left(\frac{j \gamma_{th}}{\bar{\gamma} \Omega_g} + \frac{1}{\kappa_{[c]}} \right)^{-d}, \end{aligned} \quad (17)$$

$$\begin{aligned} \mathcal{F}_3 &= \sum_{i=0}^L \sum_{j=0}^M \sum_{a,b,c,d} \binom{L}{i} \binom{M}{j} (-1)^{i+j} e^{-\gamma_{th} \mathcal{A}} \\ &\quad \frac{\mathcal{X}_{a,b}(\mathcal{Q}_1) \mathcal{X}_{c,d}(\mathcal{Q}_2)}{(\mu_{[a]})^b (\kappa_{[c]})^d} \left(\frac{\gamma_{th}}{\bar{\gamma} \xi} \mathcal{B} + \frac{1}{\mu_{[a]}} \right)^{-b} \left(\frac{j \gamma_{th}}{\bar{\gamma} \Omega_g} + \frac{1}{\kappa_{[c]}} \right)^{-d}, \end{aligned} \quad (18)$$

and $\mathcal{A} = (i \Omega_g + j (\xi + 1) \Omega_h) / (\Omega_g \Omega_h \xi \bar{\gamma})$, $\mathcal{B} = (i \Omega_g + j \Omega_h) / (\Omega_g \Omega_h \xi \bar{\gamma})$, $\sum_{a=1}^{\alpha(U)} \sum_{b=1}^{\tau_a(U)} \sum_{c=1}^{\alpha(V)} \sum_{d=1}^{\tau_c(V)}$ $= \sum_{a,b,c,d}$

Proof: Using the cdf and pdf of AS which are given below and following the same process provided in Appendix, we will obtain Eq. (15).

$$F_X(x) = \left(1 - e^{-x/\bar{x}} \right)^K \quad (19)$$

$$f_X(x) = \frac{K}{\bar{x}} \sum_{i=0}^{K-1} \binom{K}{i} (-1)^i e^{-x(i+1)/\bar{x}} \quad (20)$$

where $\bar{x} = \mathbb{E}[x]$ and K is number of antennas.

IV. ASYMPTOTIC ANALYSIS

The exact expression of OP obtained in previous section is too complicated to make the relationship between key system parameters and OP. To gain better insights, here we present asymptotic expression by applying the first order Taylor series expansion $e^{-x} \approx (1 - x)$ and using [12, Eq. (8.354.1)] into (14) and (15) and by neglecting the higher order terms we get

$$\mathcal{P}_{out,\tau}^{\infty} \underset{\bar{\gamma} \rightarrow \infty}{\approx} \mathcal{G}_d^\tau \left(\frac{\gamma_{th}}{\bar{\gamma}} \right)^{\mathcal{G}_d^\tau}, \quad (21)$$

where diversity order $\mathcal{G}_d^\tau = \min(L, M)$ and coding gain are respectively

$$\mathcal{G}_c^\tau = \begin{cases} \mathcal{G}_{c_1}^\tau, & L < M \\ \mathcal{G}_{c_2}^\tau, & M < L \\ \mathcal{G}_{c_1}^\tau + \mathcal{G}_{c_2}^\tau, & L = M, \end{cases} \quad (22)$$

with $\tau \in \{BF, AS\}$, the respective values of coding gain for beamforming and antenna selection are provided as follows:

$$\mathcal{G}_{c_1}^{BF} = \left[\frac{1}{\xi \Omega_h} \right]^L \sum_{i=0}^L \sum_{a=1}^{\alpha(U)} \sum_{b=1}^{\tau_a(U)} \binom{L}{i} \mathcal{X}_{a,b}(\mathcal{Q}_1) \frac{\Gamma(i+b)(\mu_{[a]})^i}{\Gamma(N_1)\Gamma(L+1)}, \quad (23)$$

$$\begin{aligned} \mathcal{G}_{c_2}^{BF} &= \left[\frac{1}{\Omega_g \xi} \right]^M \sum_{i=0}^M \sum_{j=0}^i \sum_{k=0}^{M-i} \sum_{a,b,c,d} \binom{M}{i} \binom{i}{j} \binom{M-i}{k} \\ &\times \frac{\xi^i \mathcal{X}_{a,b}(\mathcal{Q}_1) \mathcal{X}_{c,d}(\mathcal{Q}_2) \Gamma(k+b)\Gamma(j+d)}{(\mu_{[a]})^{-k} (\kappa_{[c]})^{-j} \Gamma(b)\Gamma(d)\Gamma(M+1)}, \end{aligned} \quad (24)$$

$$\mathcal{G}_{c_1}^{AS} = \left[\frac{1}{\xi \Omega_h} \right]^L \sum_{i=0}^L \sum_{a=1}^{\alpha(U)} \sum_{b=1}^{\tau_a(U)} \binom{L}{i} \mathcal{X}_{a,b}(\mathcal{Q}_1) \frac{\Gamma(i+b)}{\Gamma(b)} (\mu_{[a]})^i, \quad (25)$$

$$\begin{aligned} \mathcal{G}_{c_2}^{AS} &= \left[\frac{1}{\xi \Omega_g} \right]^M \sum_{j=0}^M \sum_{k=0}^j \sum_{a,b,c,d} \binom{M}{j} \binom{j}{k} \xi^k (\xi+1)^{M-j} \\ &\times \frac{\mathcal{X}_{a,b}(\mathcal{Q}_1) \mathcal{X}_{c,d}(\mathcal{Q}_2) \Gamma(j+b-k)\Gamma(k+d)}{(\mu_{[a]})^{k-j} (\kappa_{[c]})^{-k} \Gamma(b)\Gamma(d)}. \end{aligned} \quad (26)$$

A. Special cases

Above derived closed-form expressions are easy to compute but it doesn't provide useful insight into the system parameters, i.e., number of antennas, number of interferer and respective powers. Thus, we examine the special case for equal interference power, where, $P_{1,j} \triangleq P_1^I$, $\Omega_{1,j} \triangleq \Omega_1$, $P_{R,i} \triangleq P_R^I$ and $\Omega_{R,i} \triangleq \Omega_R$. Therefore, respective coding gain for each transmission scheme are as follows:

$$\mathcal{G}_{c_1}^{BF} = \frac{1}{\Gamma(L+1)} \left[\frac{1}{\xi \Omega_h} \right]^L \sum_{i=0}^L \binom{L}{i} \frac{\Gamma(i+N_1)}{\Gamma(N_1)} \left(\frac{1}{P_1^I \Omega_1} \right)^{-i}, \quad (27)$$

$$\begin{aligned} \mathcal{G}_{c_2}^{BF} &= \left[\frac{1}{\xi \Omega_g} \right]^M \sum_{i=0}^M \sum_{j=0}^i \sum_{k=0}^{M-i} \binom{M}{i} \binom{i}{j} \binom{M-i}{k} \\ &\times \frac{\xi^i \Gamma(k+N_1)\Gamma(j+N_R)}{\Gamma(M+1)(P_1^I \Omega_1)^{-k}(P_R^I \Omega_R)^{-j}}, \end{aligned} \quad (28)$$

$$\mathcal{G}_{c_1}^{AS} = \left[\frac{1}{\xi \Omega_h} \right]^L \sum_{i=0}^L \binom{L}{i} \frac{\Gamma(i+N_1)}{\Gamma(N_1)(P_1^I \Omega_1)^{-i}}, \quad (29)$$

$$\begin{aligned} \mathcal{G}_{c_2}^{AS} &= \left[\frac{1}{\xi \Omega_g} \right]^M \sum_{j=0}^M \sum_{k=0}^j \binom{M}{j} \binom{j}{k} \\ &\times \frac{(\xi+1)^{M-j} \xi^k \Gamma(j+N_1-k)\Gamma(k+N_R)}{\Gamma(N_1)\Gamma(N_R)(P_1^I \Omega_1)^{k-j}(P_R^I \Omega_R)^{-k}}. \end{aligned} \quad (30)$$

Remark (Assume $P_1^I = P_2^I = P_R^I = P_I$): From the above expressions, we can deduce that the achievable diversity order depends on the interference power level. If interference power remains fixed on condition $P_I \ll P$, then the achievable diversity order of the proposed system is $\min(L, M)$. However, when P_I increases to the same level of P , such that $\frac{P}{P_I}$ remains constant, then the diversity order is reduced to zero, which is in consent with [3].

V. OPTIMIZATION OF POWER ALLOCATION

In this section, we analyze the optimization of power allocation to minimizing the OP of our considered system in the presence of CCI. The problem of power allocation optimization can be formulated for total power constraint $P_T = P_1 + P_2 + P_R$ of given system as

$$\begin{aligned} P_1^*, P_2^*, P_R^* &= \arg \min_{P_1, P_2, P_R} \mathcal{P}_{\text{out}}^\infty(\gamma_{th}) \\ \text{subject to } P_1 + P_2 + P_R &\leq P_T, P_1, P_2, P_R > 0, \end{aligned} \quad (31)$$

where P_T is the total transmit power. For our analysis, we have assumed $P_1 = P_2 = P$. Now, the optimized power for respective transmission scheme are derived as follow:

A. Beamforming

The asymptotic expression given in (21) can be represented as

$$\begin{aligned} \mathcal{P}_{\text{out}, BF}^\infty &= \mathcal{S}_{BF} \left[\frac{\gamma_{th} N_0}{\Omega_h} \right]^L \frac{1}{(P_T - 2P)^L} \\ &+ \left[\frac{\gamma_{th} N_0}{\Omega_g} \right]^M \sum_{i=0}^{m_g M} \mathcal{T}_{BF}(i) \frac{1}{(P_T - 2P)^{m_g M-i} P_i}, \end{aligned} \quad (32)$$

where $\mathcal{S}_{BF} = \sum_{i=0}^L C_i^L \Gamma(i+N_1)(\gamma_{th} N_0)^L (P_1^I \Omega_1)^i / \Gamma(L+1)\Gamma(N_1)$ and $\mathcal{T}_{BF}(i) = \sum_{j=0}^i \sum_{k=0}^{M-i} C_i^M C_j^k C_k^{M-1} \Gamma(k+N_1)\Gamma(j+N_R)(P_1^I \Omega_1)^k (P_R^I \Omega_R)^j / \Gamma(M+1)$.

Now performing the second derivative test w.r.t. P , we deduce that $\mathcal{P}_{\text{out}, BF}^\infty$ is strictly convex function of P in the range $P \in (0, P_T/2)$. Thus, by equating its first derivative of (32) w.r.t. P to zero, optimal value of P can be determined by solving

$$\begin{aligned} \mathcal{S}_{BF} \left[\frac{\gamma_{th} N_0}{\Omega_h} \right]^L \frac{2L}{(P_T - 2P)^{L+1}} + \left[\frac{\gamma_{th} N_0}{\Omega_g} \right]^M \sum_{i=0}^M \mathcal{T}_{BF}(i) \\ \times \left(\frac{2(M-i)}{(P_T - 2P_S)^{M-i+1} P_S^i} - \frac{i}{(P_T - 2P_S)^{M-i} P_S^{i+1}} \right) = 0. \end{aligned} \quad (33)$$

As a special case for $L = M = 1$, from the above equation we will obtain the optimum value of P as

$$P^* = \frac{P_T \sqrt{\Omega_h}}{2\sqrt{\Omega_h} + \sqrt{2[P_R^I \Omega_R N_R \Omega_h \Omega_g + (\Omega_h + \Omega_g)(1 + P_1^I \Omega_1 N_1)]}}. \quad (34)$$

Thus, the optimum relay power can be expressed as $P_R^* = P_T - 2P^*$. For other values of L, M the optimum values of P and P_R can be obtained by using Numerical method.

B. Antenna Selection

Similarly, the asymptotic OP for AS scheme using (21), (29) and (30) can be expressed as

$$\begin{aligned} \mathcal{P}_{\text{out}, AS}^\infty &= \mathcal{S}_{AS} \left[\frac{\gamma_{th} N_0}{\Omega_h} \right]^L \frac{1}{(P_T - 2P)^L} \\ &+ \left[\frac{\gamma_{th} N_0}{\Omega_g} \right]^M \sum_{j=0}^M \sum_{k=0}^j \mathcal{T}_{AS}(j,k) \frac{1}{(P_T - 2P)^{M-k} P^{-k}}, \end{aligned} \quad (35)$$

where $\mathcal{S}_{AS} = \sum_{i=0}^L C_i^L \Gamma(i+N_1)(P_1^I \Omega_1)^i / \Gamma(N_1)$ and $\mathcal{T}_{AS}(j,k) = C_j^M C_k^j \Gamma(j+N_1-k)\Gamma(k+N_R)(P_1^I \Omega_1)^{j-k} (P_R^I \Omega_R)^k / \Gamma(N_1)\Gamma(N_R)$.

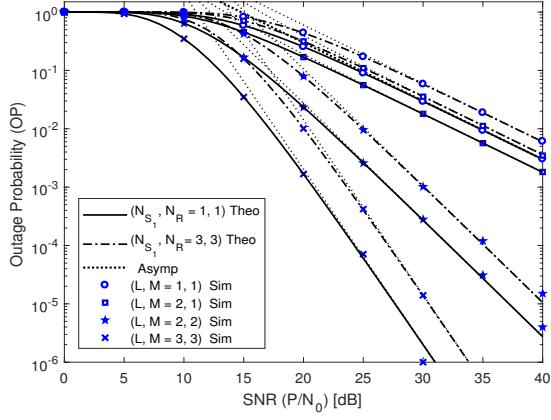


Fig. 2. OP vs SNR for BF Multi-antenna BRN system.

By following the same procedure as beamforming, we can show that $\mathcal{P}_{out,AS}^\infty$ is a strictly convex function of P in the range $P \in (0, P_T/2)$. Therefore, the optimal relay position can be obtained from

$$\begin{aligned} S_{AS} \left[\frac{\gamma_{th} N_0}{\Omega_h} \right]^L \frac{2L}{(P_T - 2P)^{L+1}} + \left[\frac{\gamma_{th} N_0}{\Omega_g} \right]^M \sum_{j=0}^M \sum_{k=0}^j \mathcal{T}_{AS}(j, k) \\ \times \left(\frac{kP^{k-1}}{(P_T - 2P)^{M-k}} - \frac{2(k-M)P^k}{(P_T - 2P)^{M-k+1}} \right) = 0. \quad (36) \end{aligned}$$

For $L = M = 1$, the above equation can be solved to obtain optimum power as

$$P^* = \frac{P_T \sqrt{\Omega_h (1 + P_R^I \Omega_R N_R)}}{2 \sqrt{\Omega_h (1 + P_R^I \Omega_R N_R)} + \sqrt{2 (1 + P_1^I \Omega_1 N_1) (\Omega_h + \Omega_g)}}. \quad (37)$$

Similarly, the optimum relay power can be expressed as $P_R^* = P_T - 2P^*$. Numerical methods can be used to obtain the solution of optimum power for higher values of L and M .

VI. NUMERICAL AND SIMULATION RESULTS

In this section, numerical results are presented to validate our analysis through Monte-Carlo simulations. Without losing generality, we have assumed $\gamma_{th}=3$ dB, all the average channel gains and noise variances to be unity. Simulations are averaged over 1 million iterations.

Fig. 2 and 3 illustrates the OP expressions (given in (14) and (15)) versus (vs) transmit signal-to-noise ratio (SNR), $\bar{\gamma} = P/N_0$ of BF and AS of the interference limited multiantenna BRN respectively. Here we have assumed $P_1^I = P_R^I = P_2^I = 20$ dB. It is clear that the simulation result having a good match with the analytical curve shows the validity of our analysis. We have shown two sets of simulation results in each figure, one with a different number of interferer and another with varying the number of antennas on each source terminal. We observed that the system OP decreases with increasing L or M and increases with increasing interferer.

Fig. 4 compares the two transmission schemes based on their OP and simulation result has the good match with analytical results. It is evident that the BF is outperforming

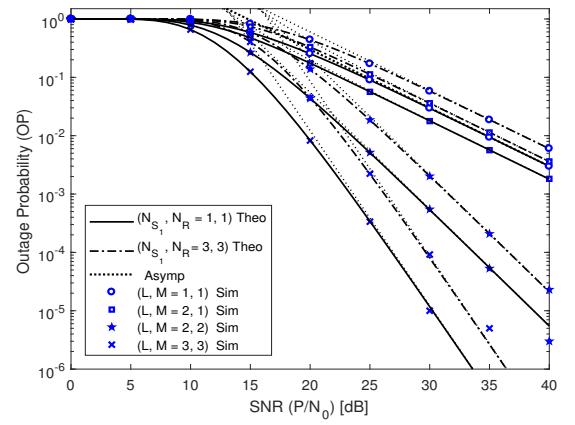


Fig. 3. OP vs SNR for AS Multi-antenna BRN system.

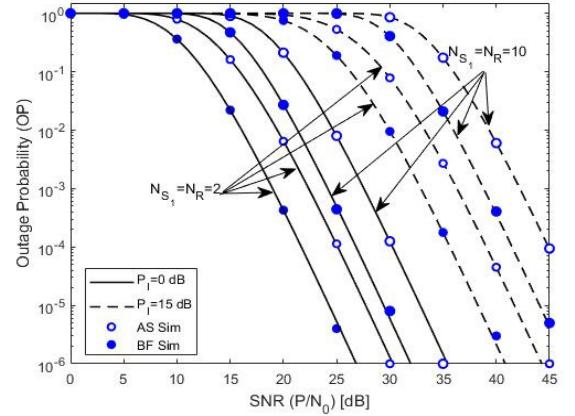


Fig. 4. OP vs SNR of Comparision of AS and BF Multi-antenna BRN system for $L = M = 3$.

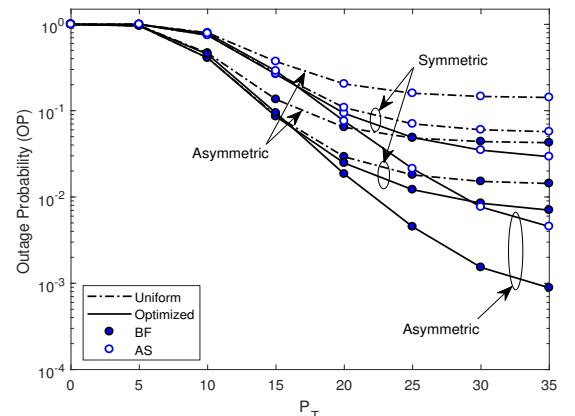


Fig. 5. OP vs P_T for optimization of power allocation for AS & BF multi-antenna BRN system for $L = M = 3$, $N_1 = N_R = 3$.

the AS and both shows the same performance degradation in the presence of interference.

In Fig. 5, we illustrate the comparison of uniform and optimized power allocation for both the transmission scheme. In uniform power allocation, power is distributed as $P_R = 2P = P_T/2$. Here we assume that the ratio of transmitted power to interference is kept constant. It is further divided into two categories as symmetric and asymmetric interference power profile. For symmetric case, $P - P_1^I = 15\text{dB}$, $P_R - P_R^I = 15\text{dB}$, while for asymmetric $P - P_1^I = 0\text{dB}$, $P_R - P_R^I = 30\text{dB}$. In the case of asymmetric interference power profile, in comparison to the non-optimized case, optimized case decreases the OP by more than ten orders of magnitude for both the transmission schemes. It is visible that the asymmetric case outperforms the symmetric one. As a concluding point, the effect of optimization is visible only after 10 dB of total power, thus for low SNR regime the optimum choice of power is $P_R = 2P = P_T/2$.

VII. CONCLUSIONS

In this paper, we proposed and analyzed the performance of multiantenna BRN for both BF and AS scheme in the presence of CCI. We have derived the tight lower bound analytical expression for the OP of both BF and AS schemes for arbitrary SINR, which gives efficient and fast mean to evaluate the system performance. Moreover, informative and simplified asymptotic outage expressions were deduced. This enable the characterization of key system parameters such as the number of source antennas (L, M), the number of interferes (N_1, N_2, N_R) and interference power on the system outage performance. Based on high SNR analysis, both schemes offer same diversity order as $\min(L, M)$. Further, we investigate the optimization problem of power allocation between source and relay terminal to minimize the outage. Herein, we declare that our results are in consent and generalize the previously reported results.

APPENDIX

Equation (7) can be written as

$$\gamma_{S_\rho \rightarrow S_\omega}^{BF} \leq \left(\gamma_{S_\rho \rightarrow S_\omega}^{BF} \right)^{up} = \min \left(\frac{\xi \gamma_\omega}{\gamma_\lambda}, \frac{\xi \gamma_\rho}{(\xi \gamma_R + \gamma_\lambda)} \right) \quad (38)$$

Now let us consider $(\rho, \omega, \lambda) \in (2, 1, \delta)$. We know that OP can be written as function of $\gamma_{S_2 \rightarrow S_1}^{BF}$ as

$$\begin{aligned} \mathcal{P}_{out}^{BF} &= F_{\left(\gamma_{S_2 \rightarrow S_1}^{BF} \right)^{up}} (\gamma_{th}) = \Pr \left(\left(\gamma_{S_2 \rightarrow S_1}^{BF} \right)^{up} \leq \gamma_{th} \right) \\ &= 1 - \Pr \left(\frac{\xi \gamma_1}{\gamma_\delta} > \gamma_{th} \right) \Pr \left(\frac{\xi \gamma_2}{(\xi \gamma_R + \gamma_\delta)} > \gamma_{th} \right) \\ &= 1 - \left(1 - \Pr \left(\|\mathbf{h}^2\| \leq \frac{\gamma_\delta \gamma_{th}}{\xi \bar{\gamma}} \right) \right) \\ &\quad \times \left(1 - \Pr \left(\|\mathbf{g}^2\| \leq \frac{(\xi \gamma_R + \gamma_\delta) \gamma_{th}}{\xi \bar{\gamma}} \right) \right) \\ &= 1 - \frac{1}{\Gamma(L)\Gamma(M)} \int_{w=0}^{\infty} \int_{v=0}^{\infty} \Gamma \left(L, \frac{(w+1)\gamma_{th}}{\Omega_h \xi \bar{\gamma}} \right) \\ &\quad \times \Gamma \left(M, \frac{\gamma_{th}(\xi(v+1)+(w+1))}{\Omega_g \xi \bar{\gamma}} \right) f_W(w) f_V(v) dw dv. \end{aligned} \quad (39)$$

The pdf of W is given as [7]

$$f_W(w) = \sum_{a=1}^{\alpha(\mathcal{Q}_1)} \sum_{b=1}^{\tau_a(\mathcal{Q}_1)} \mathcal{X}_{a,b}(\mathcal{Q}_1) \frac{(\mu_{[a]})^{-b}}{\Gamma(b)} w^{b-1} e^{-\frac{w}{\mu_{[a]}}}, \quad (40)$$

where $\mathcal{Q}_1 = diag(\mu_1, \mu_2, \dots, \mu_{N_1})$, $\alpha(\mathcal{Q}_1)$ is defined as number of distinct diagonal element of \mathcal{Q}_1 , $\mu_{[1]} > \mu_{[2]} > \dots > \mu_{[\alpha(\mathcal{Q}_1)]}$, $\mu_{[a]} = \frac{P_{1,a}\Omega_{1,a}}{N_0}$, $\mathbb{E}[f_{1,a}] = \Omega_{1,a}$, $\tau_a(\mathcal{Q}_1)$ is the multiplicity of $\mu_{[a]}$, $\mathcal{X}_{a,b}(\mathcal{Q}_1)$ is the $(a, b)_{th}$ characteristic coefficient of \mathcal{Q}_1 . Similarly the pdf of V can be expressed as

$$f_V(v) = \sum_{c=1}^{\alpha(\mathcal{Q}_2)} \sum_{d=1}^{\tau_c(\mathcal{Q}_2)} \mathcal{X}_{c,d}(\mathcal{Q}_2) \frac{(\kappa_{[c]})^{-d}}{\Gamma(d)} v^{d-1} e^{-\frac{v}{\kappa_{[c]}}}, \quad (41)$$

where $\mathcal{Q}_2 = diag(\kappa_1, \kappa_2, \dots, \kappa_{N_r})$, $\kappa_{[c]} = \frac{P_{R,c}\Omega_{R,c}}{N_0}$, $\mathbb{E}[f_{R,c}] = \Omega_{R,c}$, $\alpha(\mathcal{Q}_2)$, $\mathcal{X}_{c,d}(\mathcal{Q}_2)$, $\tau_c(\mathcal{Q}_2)$ are defined similarly as (40). Now by using the fact $\Gamma(n, x) = (n-1)!e^{-x} \sum_{m=0}^{n-1} \frac{x^m}{m!}$ and after some mathematical manipulation we will get (14).

REFERENCES

- [1] R. H. Louie, Y. Li, and B. Vucetic, "Practical physical layer network coding for two-way relay channels: performance analysis and comparison," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 764–777, Feb. 2010.
- [2] N. Yang, P. L. Yeoh, M. Elkashlan, I. B. Collings, and Z. Chen, "Two-way relaying with multi-antenna sources: Beamforming and antenna selection," *IEEE Trans. Veh. Technol.*, vol. 61, no. 9, pp. 3996–4008, 2012.
- [3] E. Soleimani-Nasab, M. Matthaiou, M. Ardebilipour, and G. K. Karagiannidis, "Two-way af relaying in the presence of co-channel interference," *IEEE Trans. Commun.*, vol. 61, no. 8, pp. 3156–3169, 2013.
- [4] S. S. Ikki and S. Aissa, "Performance analysis of two-way amplify-and-forward relaying in the presence of co-channel interferences," *IEEE Trans. Commun.*, vol. 60, no. 4, pp. 933–939, 2012.
- [5] S. Hatamnia, S. Vahidian, S. Aissa, B. Champagne, and M. Ahmadian-Attari, "Network-coded two-way relaying in spectrum-sharing systems with quality-of-service requirements," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1299–1312, 2017.
- [6] T. T. Duy, T. Q. Duong, D. B. da Costa, V. N. Q. Bao, and M. Elkashlan, "Proactive relay selection with joint impact of hardware impairment and co-channel interference," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1594–1606, Jan. 2015.
- [7] G. Zhu, C. Zhong, H. A. Suraweera, Z. Zhang, and C. Yuen, "Outage probability of dual-hop multiple antenna af systems with linear processing in the presence of co-channel interference," *IEEE Trans. Wireless Commun.*, 2014.
- [8] M. Li, M. Lin, W.-P. Zhu, Y. Huang, K.-K. Wong, and Q. Yu, "Performance analysis of dual-hop mimo af relaying network with multiple interferences," *IEEE Trans. on Veh. Technol.*, vol. 66, no. 2, pp. 1891–1897, 2017.
- [9] M. Li, L. Bai, Q. Yu, and J. Choi, "Optimal beamforming for dual-hop MIMO AF relay networks with cochannel interferences," *IEEE Trans. Signal Processing*, vol. 65, no. 7, pp. 1825–1840, 2017.
- [10] H. Phan, F.-C. Zheng, and T. M. C. Chu, "Physical-layer network coding with multiantenna transceivers in interference limited environments," *IET Commun.*, vol. 10, no. 4, pp. 363–371, 2016.
- [11] I. Khan, K. K. Dhulipudi, and P. Singh, "Performance analysis of multiantenna two-way relay network with co-channel interference," in *Proc. IEEE ANTS*. IEEE, 2017, pp. 1–6.
- [12] I. S. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic Press, 2007.
- [13] H. Ding, C. He, and L. Jiang, "Performance analysis of fixed gain MIMO relay systems in the presence of co-channel interference," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 1133–1136, 2012.

Outage Analysis of an Asymmetric dual hop PLC-VLC system for Indoor Broadcasting

Manan Jani*, Parul Garg†, and Akash Gupta‡

Division of Electronics and Communication Engineering

Netaji Subhas University of Technology

Dwarka, New Delhi-110078, India

E-mail:mananjani@outlook.com*

parul_saini@yahoo.co.in†

akashgemiini@gmail.com‡

Abstract—We propose and investigate the performance of a novel asymmetric dual hop relay based power line communication (PLC) and visible light communication (VLC) system for the purpose of indoor broadcasting. The PLC link experiences log-normal fading and additive noises. The PLC link is used as a back-haul link for the VLC downlink transmission system which serve multiple end users. The characteristics of the VLC links are dependent on the random position of the end users. Novel closed form expressions for the cumulative distribution function (CDF) of the end to end signal to noise ratio (SNR) of the proposed system is derived. Also, capitalising on these derived entities, the system's performance is evaluated in terms of the outage probability (OP) metric.

Index Terms—power line communication, visible light communication, decode and forward (DF) protocol, outage probability

I. INTRODUCTION

The growing global usage of wireless information and data access has lead to the congestion of the radio frequency (RF) spectrum [1]. Therefore, the researchers and the industries are shifting to the optical wireless communication (OWC) systems because they offer un-licensed and huge available bandwidth. More specifically, the visible light communications (VLC) systems have emerged as an attractive high data rate transmission OWC technology. They make use of the visible optical spectrum and utilise the light emitting diodes (LEDs) for the purpose of illuminating the indoor areas and for the transmission of information signals as well. In conjunction with the availability of abundant spectrum, VLC systems possess other multiple offerings such as ease of accessibility, high capacity, easy deployment, transmission security etc. However, for the purpose of communication, the indoor VLC systems must be attached to a back-bone base network or station [2].

The most efficient and economical option to resolve this issue of the VLC systems is the deployment of power line communication (PLC) system as the back-bone network to connect the indoor VLC systems to outdoor base stations. Because of their huge availability, existing infrastructure and their ability to transmit information as well as power to drive

the LEDs, PLC systems act as a potential backhaul solution to the VLC systems [2].

The integration of VLC systems with the power cables was first studied experimentally by the researchers in [3], where they explored the performance of a low data rate PLC-VLC system utilising single carrier (SC) binary phase shift keying (BPSK) digital carrier scheme. The experimental setup was designed utilising existing narrow-band power cables as a source of the energy to the white LEDs. Other digital modulation keying schemes like on-off keying (OOK), orthogonal frequency division multiplexing (OFDM) schemes etc. were employed to study the performance of the heterogeneous PLC-VLC systems in [4], [5].

An indoor family broadcasting PLC-VLC system was proposed by the authors in [6], where they presented a two-lamp experimental prototype and investigated the performance in the measures of maximum achievable rate. In [7], the authors studied a decode and forward (DF) based PLC-VLC duplex framework system. They proposed a multi-user broadband system to enhance the quality-of-service (QoS) of the integrated system. However, the channel statistics of both the PLC and VLC systems were not explored. To study the noise scenarios of these integrated hybrid systems, the authors in [8] investigated a deterministic generalised channel model. However, the stochastic parameters of the channels were not taken into account.

Most of the previous work related to the analysis and modeling of these composite systems employed deterministic channels, thus, their statistical performance had not been analysed. The authors in [9] studied and investigated the performance of a relay aided PLC-VLC indoor communication system for multimedia broadcasting. They used statistical channel modeling for the performance analysis of a single PLC-VLC link, in which a single end user in the VLC link was connected to base back-bone PLC network. However, in a practical indoor scenario, the VLC downlink systems have multiple access points (APs) which serve end user terminals (UTs) distributed randomly in the AP coverage area. To the best of the authors' knowledge, the statistical analysis of such systems employing multiple access points at the VLC end

has not been studied in the literature. Therefore, motivated by these aforementioned facts, the authors propose and investigate a stochastic PLC-VLC model for indoor communication purpose. The key contributions of the study are as listed:

- 1) A deeply integrated dual hop PLC-VLC system for indoor multimedia broadcasting purpose is presented. The VLC link has multiple APs attached to the ceiling which serve randomly distributed UTs in their coverage area.
- 2) Based on the stochastic models of the PLC and the VLC link, the statistics of the end to end signal to noise ratio (SNR) of the composite system are obtained.
- 3) We investigate the performance of the hybrid system in terms of outage probability (OP) metric.
- 4) The system performance is studied for different values of impulsive noise conditions and varying indoor parameters such as room size, number of users etc.

The remainder of the study is organized as: The DF relay aided dual hop PLC-VLC system model is presented in section II, which is followed by the channel model of the system in section III. In section IV, we derive the analytical closed form expressions of the cumulative distributive function (CDF) of the end to end signal to noise ratio (SNR) of the considered PLC-VLC system. Using them, the closed form expressions of the outage probability are also derived and presented. Finally, we present the numerical results in section V. Finally, some concluding statements are presented in section VI.

II. SYSTEM MODEL

We consider an asymmetric dual hop DF relay assisted PLC-VLC system. The signal is transmitted from the base network station to the VLC link. The VLC link is connected to the backhaul PLC link with the help of a DF relay as shown in the Fig.1. The power lines which carry data from base station to the VLC link encounter several points of discontinuities, which results in multiple delayed and distorted copies of actual signal. These multiple copies result in multi-path fading effects in the PLC links [10]. In addition to fading, the PLC channels are also corrupted by additive noises. The PLC link transmits data to relay, R as

$$r_\rho = h_\rho x + n_\rho, \quad (1)$$

where x is the source signal. The channel fading of the PLC link is denoted by h_ρ and the additive noises of the PLC system are represented by n_ρ .

This received data is decoded by R which retransmits it to the VLC link on the optical visible light range. At the downlink VLC end, multiple VLC access points (APs) composed of LEDs are present. Each LED is positioned at a height L , the radius r_i and at an angle θ_i from the i^{th} end user on the polar co-ordinate axis. The APs cater the I end users at the VLC end. These I users are distributed uniformly over the coverage area of the VLC system. The received signal r_v from R can be given as

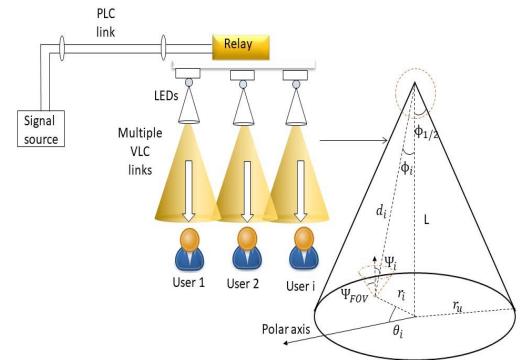


Fig. 1. System Model of the proposed PLC-VLC system

$$r_v = h_v \hat{x} + n_v, \quad (2)$$

where the decoded data is denoted by \hat{x} , h_v incorporates the fading in the VLC links and the noises in the VLC links are represented by n_v . At the user end, the photo-detectors (PD) are available which convert the received optical signal into electrical information with the help of direct detection (DD) technique [11].

III. CHANNEL MODEL

In this paper, it is assumed that the PLC link experiences log normal fading and is affected by the additive noises [9]. The downlink VLC system is modeled stochastically as a function of the uniformly distributed location of the I end users.

A. Power Line Communication link

Along the power cables, the signal from the base station passes through several discontinuous points and impedance mismatches giving rise to multiple delayed and distorted versions of the signal along with the original signal itself. This results in multi-path fading. The log normal distribution is widely accepted in the literature as the channel fading model to model the multi-path fading effects in the PLC links [9], [10]. The probability density function (PDF) of these log normal fading multipliers can be presented as

$$f_{H_\rho}(h_\rho) = \frac{1}{h_\rho \sqrt{2\pi\sigma_\rho^2}} \exp\left(-\frac{(\ln h_\rho - \mu_\rho)^2}{2\sigma_\rho^2}\right), \quad (3)$$

where μ_ρ and σ_ρ^2 denote the mean and variance of the log normally distributed random variable $\ln h_\rho$.

In contrast to the wireless communication, the power lines experience both additive background and impulsive noises. The main causes of background noise in the PLC channel are the low-power components in the system. On the other hand, impulsive noise is generated because of the random transient switching of the electrical appliances connected to the cables. To analyze the combined effects of these noises, a statistical

model was introduced in [12]. Using this model, the noise n_ρ in the PLC channel is given as

$$n_\rho = n_{b*} + n_{i*}, \quad (4)$$

where n_{b*} and n_{i*} represent the samples of background noise and impulsive noise respectively. The sample n_{b*} is modeled as zero mean additive white gaussian noise (AWGN) process with standard deviation σ_{BN} , whereas, the sample n_{i*} is characterized as a randomly distributed Poisson process, which can be given as

$$n_{i*} = n_{P*} \times n_{I*}, \quad (5)$$

where n_{P*} models the transient arrivals of the random impulsive noise in the PLC link and n_{I*} is zero mean AWGN process with standard deviation σ_{IN} . It is assumed that n_{b*} and n_{i*} are mutually independent of each other.

Using [13, Eq.4] and considering only the real part of the noise, we get the PDF of n_ρ as

$$f_{N_\rho}(n_\rho) = (1 - P_I) \left[\frac{1}{\sqrt{2\pi\sigma_{BN}^2}} \exp\left(-\frac{n_\rho^2}{2\sigma_{BN}^2}\right) \right] + (P_I) \left[\frac{1}{\sqrt{2\pi(\sigma_{BN}^2 + \sigma_{IN}^2)}} \exp\left(-\frac{n_\rho^2}{2(\sigma_{BN}^2 + \sigma_{IN}^2)}\right) \right], \quad (6)$$

where P_I is defined as the probability of the arrival of impulsive noise. When only background noise is present in the link, the total noise power is σ_{BN}^2 , but when both background noise and impulsive noise are present in the system, the total noise power amounts to $\sigma_{BN}^2 + \sigma_{IN}^2$. Using [13, Eq.7], the P_I can be represented as

$$P_I = \lambda T_I, \quad (7)$$

where λ is the parameter that characterizes the arrival rate of the impulsive noise. Each impulsive noise fluctuation occurs for a total time T_I in the system.

Let the instantaneous SNR, γ_ρ received at R be given as

$$\gamma_\rho = \begin{cases} \gamma_{\rho_1} = \frac{E_b |h_\rho|^2}{\sigma_{BN}^2} & \text{When only background noise is present,} \\ \gamma_{\rho_2} = \frac{E_b |h_\rho|^2}{\sigma_{BN}^2 + \sigma_{IN}^2} & \text{When impulsive and background noises are present,} \end{cases} \quad (8)$$

where E_b represents the energy of the signal.

Using (3), (6), (8) and after doing some re-arrangements, the PDF of γ_ρ can be obtained as

$$f_{\Gamma_\rho}(\gamma) = (1 - \lambda T_I) \left[\frac{1}{\sqrt{2\pi\sigma_x^2\gamma}} \exp\left(-\frac{(\ln\gamma - \mu_x)^2}{2\sigma_x^2}\right) \right] + (\lambda T_I) \left[\frac{1}{\sqrt{2\pi\sigma_y^2\gamma}} \exp\left(-\frac{(\ln\gamma - \mu_y)^2}{2\sigma_y^2}\right) \right], \quad (9)$$

where $\mu_x = 2\mu_\rho + \ln\bar{\gamma}_{\rho_1}$, $\mu_y = 2\mu_\rho + \ln\bar{\gamma}_{\rho_2}$, $\sigma_x = \sigma_y = 2\mu_\rho$. Here, $\bar{\gamma}_{\rho_1} = \frac{E_b}{\sigma_{BN}^2}$ is the average SNR of the PLC channel in

the presence of background noise only and $\bar{\gamma}_{\rho_2} = \frac{E_b}{(\delta+1)\sigma_{BN}^2}$ denotes the average SNR of the channel when both background and impulsive noises are present. The δ is a parameter to represent the ratio of impulsive noise power, σ_{IN}^2 and background noise power, σ_{BN}^2 .

We use a Gamma approximation model as presented in [14] to approximate the log normal PDF with the Gamma PDF in order to obtain the closed form analytical expressions of the statistical metrics. Using this approximation, we can represent the $f_{\Gamma_\rho}(\gamma)$ as

$$f_{\Gamma_\rho}(\gamma) = (1 - \lambda T_I) \left(\frac{m_r}{\Omega_r} \right)^{m_r} \frac{\gamma^{m_r-1}}{\Gamma(m_r)} \exp\left(-\frac{\gamma m_r}{\Omega_r}\right) + (\lambda T_I) \left(\frac{m_q}{\Omega_q} \right)^{m_q} \frac{\gamma^{m_q-1}}{\Gamma(m_q)} \exp\left(-\frac{\gamma m_q}{\Omega_q}\right), \quad (10)$$

where $\Gamma(.) = \int_0^\infty e^{-z} z^{x-1} dz$ represents the Gamma function. The m_r , m_q , Ω_r and Ω_q represent the parameters of the approximated Gamma PDF which are obtained by using moment matching (MM) method [14], and can be given as $m_r = \frac{1}{(e^{\sigma_x^2}-1)}$, $m_q = \frac{1}{(e^{\sigma_y^2}-1)}$, $\Omega_r = e^{\mu_x} \sqrt{\left(\frac{1+m_x}{m_x}\right)}$ and $\Omega_q = e^{\mu_y} \sqrt{\left(\frac{1+m_y}{m_y}\right)}$.

B. Visible Light Communication link

The channel model of the VLC link is a function of the random distribution of the multiple end users distributed uniformly in the circular region (atto-cell) under the coverage of the LED. The PDF of the uniform distribution of the i^{th} user's position is given by [11]

$$f_{r_i}(r) = \frac{2r}{r_u^2}, \quad (11)$$

where r_u denotes the maximum radius which is covered by the LED.

The VLC links consist of both line of sight (LOS) and non line of sight (NLOS) components but the proportion of the NLOS components in the indoor broadcasting systems is about 5% of the signal received by the user as compared to their LOS counterparts [15]. Therefore, for the system analysis, it is assumed that each VLC AP serves the end user by the strongest possible LOS link for that user.

The channel gain of the LOS link between the i^{th} end user and the LED is given as

$$h_i = \frac{(m_i + 1)}{(2\pi d_i^2)} A_D R_D \cos^{m_i}(\phi_i) \cos(\psi_i) U(\psi_i) g(\psi_i), \quad (12)$$

where $m_i = \frac{-1}{\log_2(\cos(\phi_{1/2}))}$ is the Lambertian pattern of the radiation followed by the LED [11]. Here, $\phi_{1/2}$ represents the semi-angle of the LED. The area and the responsivity of the detector at the user end is represented by A_D and R_D respectively. The Euclidean distance between the i^{th} user and the LED is denoted by d_i . $U(\psi_i)$ and $g(\psi_i)$ represent the gains of the optical filter and the optical concentrator respectively.

We can represent the $g(\psi_i)$ in terms of the field of view (FOV), Ψ_{FOV} of the detector as

$$g(\psi_i) = \begin{cases} \frac{n_{r_i}^2}{\sin^2(\Psi_{FOV})} & 0 \leq \psi_i \leq \Psi_{FOV}, \\ 0 & \psi_i > \Psi_{FOV}, \end{cases} \quad (13)$$

where the refractive index of the optical concentrator is represented by n_{r_i} [11].

Substituting the values of the $\cos(\phi_i) = \cos(\psi_i) = L/d_i$ and $d_i = (L^2 + r_i^2)^{\frac{1}{2}}$ from the i^{th} user in (12), we obtain h_i as [16]

$$h_i = \frac{\xi(m_i + 1)L^{m_i+1}}{(r_i^2 + L^2)^{\frac{m_i+3}{2}}}, \quad (14)$$

where $\xi = \frac{A_D R_D U(\psi_i) g(\psi_i)}{2\pi}$ denotes a system dependent constant.

The PDF of the channel gain can be obtained by using (11), (14) and [17], and can be given as

$$f_{h_i}(h) = \frac{2\xi^{\frac{2}{m_i+3}} ((m_i + 1)L^{m_i+1})^{\frac{2}{m_i+3}} h^{-\frac{2}{m_i+3}-1}}{(m_i + 3)r_u^2}. \quad (15)$$

The instantaneous SNR γ_v of the VLC link can be written as

$$\gamma_v = h_i^2 \bar{\gamma}_v, \quad (16)$$

where $\bar{\gamma}_v = \frac{\varrho^2 P_{optimal}^2}{B_i N_o}$ is the average SNR of the VLC system. The parameter ϱ represents the electrical to optical conversion efficiency, $P_{optimal}$ is the transmitted optical power, B_i denotes the baseband modulation bandwidth and N_o is the spectral density of the noise.

Utilising (15) and (16), we can obtain the PDF, $f_{\Gamma_{v_i}}(\gamma)$ of γ_v of the i^{th} link as

$$f_{\Gamma_{v_i}}(\gamma) = \frac{\xi^{\frac{2}{m_i+3}} ((m_i + 1)L^{m_i+1})^{\frac{2}{m_i+3}} \gamma^{-\frac{m_i+4}{m_i+3}} \bar{\gamma}_{VLC}^{\frac{1}{m_i+3}}}{(m_i + 3)r_u^2}, \quad (17)$$

for $\gamma \in [\tau_{min}, \tau_{max}]$ where $\tau_{min} = \left[\frac{(\xi(L^{m_i+1})(m_i+1))^2}{(r_u^2 + L^2)^{m_i+3}} \right] \bar{\gamma}_v$ and $\tau_{max} = \left[\frac{(\xi(L^{m_i+1})(m_i+1))^2}{(L^2)^{m_i+3}} \right] \bar{\gamma}_v$.

IV. OUTAGE ANALYSIS OF THE PLC-VLC LINK

In this section, we analyze the outage probability (OP) metric of the proposed dual hop PLC-VLC system. For the DF relay based protocol, the equivalent end to end SNR, γ_{ee} of the system is given as

$$\gamma_{ee} = \min(\gamma_\rho, \gamma_v^m), \quad (18)$$

where γ_v^m denotes the instantaneous SNR of the LOS link with the best instantaneous SNR among the numerous LOS VLC links available.

The system outage is said to have occurred when γ_{ee} of the system falls below a specified value of the threshold SNR, γ_{th} . Therefore, we can represent OP as

$$OP = Pr(\gamma_{ee} \leq \gamma_{th}). \quad (19)$$

The cumulative distribution function (CDF) of γ_{ee} is given as

$$\begin{aligned} \mathcal{F}_{\Gamma_{ee}}(\gamma) &= Pr(\gamma_{ee} < \gamma) \\ &= Pr[\min(\gamma_\rho, \gamma_v^m) < \gamma] \\ &= Pr[\gamma_\rho < \gamma] + Pr[\gamma_v^m < \gamma] \\ &\quad - Pr[\gamma_\rho < \gamma] Pr[\gamma_v^m < \gamma]. \end{aligned} \quad (20)$$

As γ_ρ and γ_v^m are independent of each other, hence, re-writing (20), we have

$$\mathcal{F}_{\Gamma_{ee}}(\gamma) = \mathcal{F}_{\Gamma_\rho}(\gamma) + \mathcal{F}_{\Gamma_v^m}(\gamma) - \mathcal{F}_{\Gamma_\rho}(\gamma) \mathcal{F}_{\Gamma_v^m}(\gamma), \quad (21)$$

where the CDFs of the instantaneous SNRs of the PLC and VLC links are given by $\mathcal{F}_{\Gamma_\rho}(\gamma)$ and $\mathcal{F}_{\Gamma_v^m}(\gamma)$ respectively.

A. CDF of PLC Link

The CDF of the PLC link can be obtained by integrating (10) and using [18, Eq.8.4.16] as $\mathcal{F}_{\Gamma_\rho}(\gamma)$ as

$$\begin{aligned} \mathcal{F}_{\Gamma_\rho}(\gamma) &= (1 - \lambda T_I) \left[\frac{1}{\Gamma(m_r)} \mathcal{G}_{1,2}^{1,1} \left(\frac{m_r \gamma}{\Omega_r} \middle| {}_{m_r,0}^1 \right) \right] \\ &\quad + (\lambda T_I) \left[\frac{1}{\Gamma(m_q)} \mathcal{G}_{1,2}^{1,1} \left(\frac{m_q \gamma}{\Omega_q} \middle| {}_{m_q,0}^1 \right) \right], \end{aligned} \quad (22)$$

where $\mathcal{G}_{k,l}^{i,j}(\cdot | \cdot)$ is the Meijer's-G function [18, Section 2.24].

B. CDF of VLC Link

To get the CDF of i^{th} VLC link, we integrate (17) within the limits $[\tau_{min}, \tau_{max}]$ and obtain the expression for $\mathcal{F}_{\Gamma_{v_i}}(\gamma)$ as

$$\begin{aligned} \mathcal{F}_{\Gamma_{v_i}}(\gamma) &= -\frac{\xi^{\frac{2}{m_i+3}} ((m_i + 1)L^{m_i+1})^{\frac{2}{m_i+3}}}{r_u^2} \left(\frac{\gamma}{\bar{\gamma}_{VLC}} \right)^{\frac{-1}{m_i+3}} \\ &\quad + \left(\frac{r_i^2 + L^2}{r_i^2} \right). \end{aligned} \quad (23)$$

At the VLC end, the APs serve the end users using multiple LOS links. The user selects that link which has the maximum instantaneous SNR at that particular time. The CDF of this selected link can be given as

$$\mathcal{F}_{\Gamma_v^m}(\gamma) = \prod_{i=1}^K \mathcal{F}_{\Gamma_{v_i}}(\gamma), \quad (24)$$

where K denotes the total number of the APs at the VLC end.

Using (23), (24) and assuming identical and independent characteristics of the statistics of the SNR for all the links of the different APs, we obtain the CDF, $\mathcal{F}_{\Gamma_v^m}(\gamma)$ as

$$\mathcal{F}_{\Gamma_v^m}(\gamma) = \sum_{i=0}^K (-1)^i \binom{K}{i} \eta^i \zeta^{K-i} \left(\frac{\gamma}{\bar{\gamma}_v} \right)^{\frac{-i}{m_i+3}}, \quad (25)$$

where $\eta = \frac{\xi^{\frac{2}{m_i+3}} ((m_i + 1)L^{m_i+1})^{\frac{2}{m_i+3}}}{r_u^2}$ and $\zeta = \left(\frac{r_u^2 + L^2}{r_u^2} \right)$.

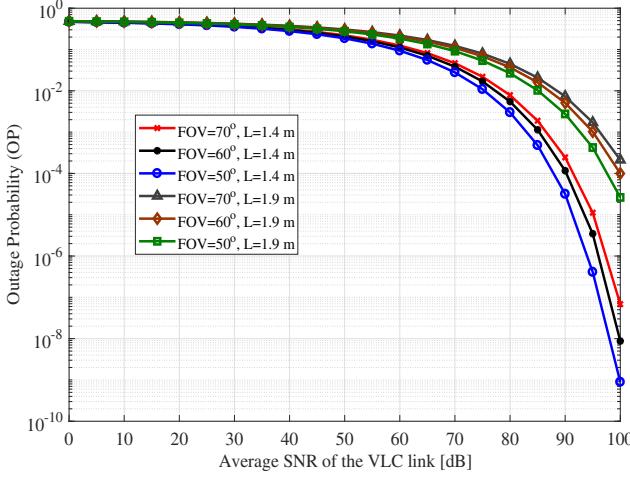


Fig. 2. Outage Probability (OP) versus $\bar{\gamma}_v$ plot for varying values of FOV and room dimensions

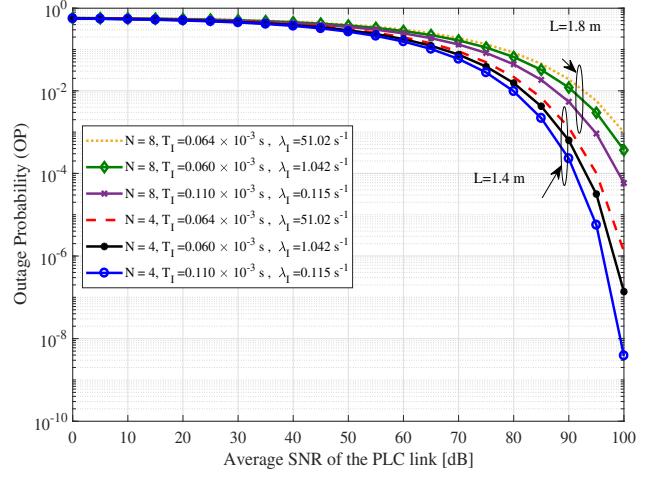


Fig. 4. Outage Probability (OP) versus $\bar{\gamma}_p$ plot for different impulsive noise distribution scenarios and varying room scenarios

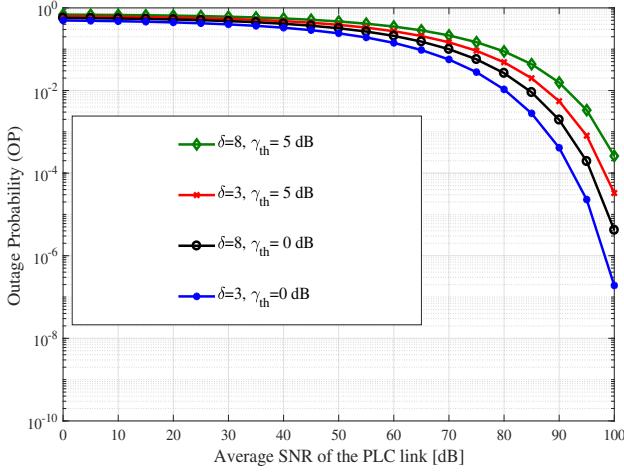


Fig. 3. Outage Probability (OP) versus $\bar{\gamma}_p$ plot for varying impulsive noise ratio, δ and different threshold SNRs, γ_{th}

C. Outage Analysis

Using (21), (22) and (25), we obtain the novel expressions of the CDF of the equivalent end to end SNR, γ_{ee} as

$$\begin{aligned} \mathcal{F}_{\Gamma_{ee}}(\gamma) &= \left[(1 - \lambda T_I) \left(\frac{1}{\Gamma(m_r)} G_{1,2}^{1,1} \left[\frac{m_r \gamma}{\Omega_r} \Big|_{m_r,0}^1 \right] \right) \right. \\ &\quad + (\lambda T_I) \left(\frac{1}{\Gamma(m_q)} G_{1,2}^{1,1} \left[\frac{m_q \gamma}{\Omega_q} \Big|_{m_q,0}^1 \right] \right) \\ &\quad \times \left(1 - \sum_{i=0}^K (-1)^i \binom{K}{i} \eta^i \zeta^{K-i} \left(\frac{\gamma}{\bar{\gamma}_v} \right)^{\frac{-i}{m_i+3}} \right) \Big] \\ &\quad + \left[\sum_{i=0}^K (-1)^i \binom{K}{i} \eta^i \zeta^{K-i} \left(\frac{\gamma}{\bar{\gamma}_v} \right)^{\frac{-i}{m_i+3}} \right]. \quad (26) \end{aligned}$$

Using (19) and (26), we get the novel closed-form expressions for the outage probability of the considered PLC-VLC

system as (see(27)).

V. NUMERICAL ANALYSIS

In this section, we investigate the performance of the proposed PLC-VLC system on the basis of the derived novel closed form expressions of outage probability (OP) metric from the earlier section. The values of m_r and m_q are taken as 1.32 [14].

Fig. 2 compares the outage performance of the system for varying values of photo-detector's FOV and different room dimensions. Here, the value of the semi-angle of the LED, $\phi_{1/2}$ is taken as 65°. The total number of VLC APs are taken as 10. The value of impulsive noise ratio parameter, δ is set at 5. It is observed that the outage performance increases when the value of the FOV of the photo-detector is decreased, which is evident as decreasing the FOV will subsequently decrease the power of the concentrator at the detector. Also, it is seen that on increasing the height of the LEDs, the outage performance degrades.

Fig. 3 presents the outage performance of the system under the variation of the impulsive noise ratio, δ and various values of threshold SNRs, γ_{th} . The $\phi_{1/2}$ is set at 65°. The total number of VLC APs are taken as 12 for the performance analysis. As evident from the plot, the system performance deteriorates considerably when the value of δ is increased. This is because the distribution of the power of the impulsive noise in the system will increase with the increase in δ . Also, it is observed that the performance of the system is enhanced when the value of γ_{th} is decreased as compared to its higher values.

In Fig. 4, the outage performance of the considered system is analysed for different impulsive noise distributions [12], varying room dimensions and different number of VLC APs. The value of the impulsive noise ratio, δ is taken as 8. The $\phi_{1/2}$ is set at 60°. The variation of the distribution of the impulsive noise is taken as high impulsive noise conditions

$$OP = \left[(1 - \lambda T_I) \left(\frac{1}{\Gamma(m_r)} G_{1,2}^{1,1} \left[\frac{m_r \gamma_{th}}{\Omega_r} \Big|_{m_r,0} \right] \right) + (\lambda T_I) \left(\frac{1}{\Gamma(m_q)} G_{1,2}^{1,1} \left[\frac{m_q \gamma_{th}}{\Omega_q} \Big|_{m_q,0} \right] \right) \right. \\ \times \left. \left(1 - \sum_{i=0}^K (-1)^i \binom{K}{i} \eta^i \zeta^{K-i} \left(\frac{\gamma_{th}}{\bar{\gamma}_v} \right)^{\frac{-i}{m_i+3}} \right) \right] + \left(\sum_{i=0}^K (-1)^i \binom{K}{i} \eta^i \zeta^{K-i} \left(\frac{\gamma_{th}}{\bar{\gamma}_v} \right)^{\frac{-i}{m_i+3}} \right). \quad (27)$$

($T_I = 0.064 \times 10^{-3} s$, $\lambda = 51.02 s^{-1}$), medium impulsive noise conditions ($T_I = 0.060 \times 10^{-3} s$, $\lambda = 1.042 s^{-1}$) and weak impulsive noise conditions ($T_I = 0.11 \times 10^{-3} s$, $\lambda = 0.115 s^{-1}$) [12]. As observed from the figure, the system performs better when the distribution of the impulsive noise is weak as compared to the highly distributed impulsive noise scenario. Also, it is seen that as the number of VLC APs increase, the system outage also increases, therefore, the overall performance of the system degrades considerably.

VI. CONCLUSION

In this paper, we propose a novel dual hop DF relay based PLC-VLC indoor broadcasting system considering multiple access points at the VLC end. Using statistical channel models, we derive novel closed form analytical expression of the cumulative distribution function (CDF) of the equivalent end to end SNR of the system. Taking this into consideration, we obtain the closed form expressions for the system outage probability. The performance of the system is observed for varying impulsive noise distributions and various indoor scenarios. It is observed from the numerical results that the system performance is improved when the distribution of the impulsive noise is weak in comparison to heavily distributed scenarios.

REFERENCES

- [1] D. Jokanovic and M. Josipovic, "RF spectrum congestion: resolving an interference case," *Microwaves, Communications, Antennas and Electronics Systems (COMCAS)*, pp. 1–4, 2011.
- [2] X. Ma et al., "Integrated power line and visible light communication system compatible with multiservice transmission," *IET Communications*, vol. 11, no. 1, pp. 104–111, 2017.
- [3] T. Komine and M. Nakagawa, "Integrated system of white LED visible-light communication and power-line communication," *IEEE Trans. Consum. Electron.*, vol. 49, no. 1, pp. 71–79, 2003.
- [4] A. Ndjiongue et al., "Low-complexity SOCPBFSK-OOK interface between PLC and VLC channels for low data rate transmission applications," *Proc. IEEE ISPLC*, pp. 226–231, March 2014.
- [5] T. Komine, S. Haruyama, and M. Nakagawa, "Performance evaluation of narrowband ofdm on integrated system of power line communication and visible light wireless communication," *International Symposium on Wireless Pervasive Computing*, pp. 6–11, 2006.
- [6] J. Song, W. Ding, F. Yang, H. Yang, B. Yu, and H. Zhang, "An indoor broadband broadcasting system based on PLC and VLC," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 299–308, June 2015.
- [7] J. Song et al., "A Cost-Effective Approach for Ubiquitous Broadband Access Based on Hybrid PLC-VLC System," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2815–2818, 2016.
- [8] S. Nlom et al., "A Simplistic Channel Model for Cascaded PLC-VLC Systems," —emphIEEE International Symposium on Power Line Communications and its Applications (ISPLC), 2017.
- [9] M. Jani, P. Garg and A. Gupta, "Modeling and Outage Analysis of DF Relay assisted mixed PLC-VLC System," *Proc. National Conference on Commun.*, Hyderabad, India, Feb. 2018.
- [10] S. Guzelgoz, H. Celebi, and H. Arslan, "Statistical characterization of the paths in multipath plc channels," *IEEE Trans. Power Del.*, vol. 26, no. 1, pp. 181–187, Jan. 2011.
- [11] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.
- [12] Y.H. Ma, P.L. So, and E. Gunawan, "Performance analysis of OFDM systems for broadband power line communications under impulsive noise and multipath effects," *IEEE Trans. Power Deliv.*, vol. 20, no. 2, pp. 674–682, 2005.
- [13] M. Jani, P. Garg and A. Bansal, "Performance analysis of a PLC system over log-normal fading channel and impulsive noise," *2015 International Conference on Computing and Network Communications (CoCoNet)*, pp. 480–484, 2015.
- [14] I. M. Kostic, "Analytical approach to performance analysis for channel subject to shadowing and fading," *IEEE Proc. Commun.*, vol. 152, no. 6, pp. 821–827, Dec. 2005.
- [15] L. Zeng, D. C. O'Brien, H. L. Minh, G. E. Faulkner, K. Lee, D. Jung, Y. Oh, and E. T. Won, "High data rate multiple input multiple output (MIMO) optical wireless communications using white LED lighting," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 9, pp. 1654–1662, Dec. 2009.
- [16] A. Gupta, N. Sharma, P. Garg, and M. S. Alouini, "Cascaded FSO-VLC communication system," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 810–813, Dec. 2017.
- [17] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. Tata McGraw-Hill Education, 2002.
- [18] A. P. Prudnikov, Y. A. Brychkov, and O. I. Marichev, *Integrals and Series. Volume 3: More Special Functions.*, 1st ed. Gordon and Breach Science Publishers, 1986.

Error performance of QAM GFDM waveform with CFO under AWGN and TWDP fading channel

Sapta Girish Neelam and P. R. Sahu

Abstract—Generalized frequency division multiplexing (GFDM) is a strong contender waveform for 5G cellular communications. This letter derives closed form symbol error rate (SER) expressions with carrier frequency offset (CFO) in the presence of additive white gaussian noise (AWGN) channel for 1. normal BPSK, QPSK and 16-QAM-GFDM and 2. Time shift and Frequency shift, offset QPSK (OQPSK) and offset 16-QAM GFDM waveforms. This letter also derives SER expressions under TWDP fading channel for BPSK and QPSK GFDM waveforms. It is observed that under AWGN channel 1. SER performance of frequency shift offset QAM (FS-OQAM) GFDM is better than conventional MF, ZF and MMSE receivers and 2. FS-OQAM-GFDM slightly outperforms time shift OQAM(TS-OQAM) GFDM at higher values of CFO. The performance analysis studies show that TWDP fading can result in a performance poorer than Rayleigh fading when two specular components which are equal in strength and antiphase cancel each other for $\bar{K} > 6$ dB. The numerically evaluated results are in close agreement with the simulated results.

Index Terms—5G, GFDM, QAM, TS-OQAM-GFDM, FS-OQAM-GFDM

I. INTRODUCTION

Waveforms under active research for the fifth generation (5G) cellular standard are either *linear* or *circular* type based on filtering process of the sub-carriers [1][2]. The linear pulse shaping waveforms under the filter bank multi carrier (FBMC) scheme are not suitable for short burst transmissions as the filter length is more than four times the symbol period and it has excessive ramp-up and ramp-down time at the start and end of the packet. The next contender, the non-orthogonal circularly filtered block based waveform, the GFDM waveform [5] preserves the advantages of OFDM while addressing its limitations such as its spectral inefficiency and drawbacks in carrier aggregation. Since GFDM is a non-orthogonal waveform, it suffers from inter-carrier interference (ICI). Either advanced receivers with high computational complexity like minimum mean square error (MMSE) or zero forcing (ZF) can be used to mitigate interference or different versions of OQAM can be used to address non-orthogonal conditions and singularity problems with the modulation matrix when even number of subsymbols are transmitted.

Time shift offset QAM (TS-OQAM) leads to overlapping of subsymbols because of time shift between two real data sequences which means that entire GFDM frame is required to recover the data because subsymbol by subsymbol detection is

The authors are with the School of Electrical Sciences, IIT Bhubaneswar, India-752050, Emails: sgn10@iitbbs.ac.in (Sapta Girish Neelam), prs@iitbbs.ac.in (P. R. Sahu)

not possible. By using unitary transformation and exploring the duality of time-frequency domain, frequency shift offset QAM (FS-OQAM) GFDM [6] is designed. A half Nyquist pulse of two subsymbols width requires a frequency shift of one half of subcarrier. FS-OQAM-GFDM can be used in low latency applications like machine to machine (*M2M*) communications and tactile internet because subsymbol by subsymbol detection is possible.

Lot of work is done on understanding the channel for millimeter-wave (mmWave) bands. Two wave with diffused power (TWDP) fading channel is a special case of fluctuating two ray (FTR) fading channel which is used to model the channel at mmWave frequencies. TWDP fading channel includes Rayleigh and Rician fading models as special cases. In this letter we derive closed form SER expressions for normal QAM GFDM, TS-OQAM-GFDM and FS-OQAM-GFDM with CFO in the presence of AWGN. In particular, closed form SER expressions for BPSK, QPSK and 16 QAM GFDM with CFO under AWGN are derived. In this letter we also derive SER expressions for normal BPSK GFDM and QPSK and OQPSK-GFDM with CFO in the presence of TWDP fading channel. Section II introduces GFDM system and Sections III and IV derives the SER expressions under AWGN channel and TWDP fading channel. Numerical analysis is done in section V and finally it is concluded in section VI.

II. SYSTEM AND CHANNEL MODEL

GFDM is a circularly filtered non-orthogonal multicarrier waveform with K subcarriers and M subsymbols in each block where $N = MK$ represents total number of GFDM symbols. The prototype filter $g_T[n]$ is circularly shifted in both time and frequency domain and can be expressed as

$$g_{k,m}[n] = g_T[\lfloor (n - mK) \rfloor_N] \exp\left(j2\pi n \frac{k}{K}\right) \quad (1)$$

where $n = 0, 1, \dots, N - 1$ is the time index. The operator $[A]_B$ represents mathematical operation A modulo B and $d_{k,m} = d_{k,m}^{(i)} + jd_{k,m}^{(q)}$ is the complex data symbol transmitted on subcarrier k and subsymbol m . GFDM signal is given as $x[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{k,m} g_{k,m}[n]$ which is also written as

$$x[n] = \sum_{l=0}^{N-1} \mathbf{d}[l] g_T[\lfloor n - \lfloor l \rfloor MK \rfloor_N] \exp\left(j\frac{2\pi n}{N}(l - \lfloor l \rfloor_M)\right) \quad (2)$$

where \mathbf{d} is a column vector which holds all the values of $d_{k,m}$. The GFDM signal is passed through an AWGN channel $z[n]$ which is Gaussian distributed of 0 mean and variance σ_n^2 as

$$y[n] = \beta x[n] + z[n] \quad (3)$$

β is the complex fading coefficient and $\beta = 1$ is assumed for AWGN channel. Assuming ideal channel equalization, after removing CP the estimate $\hat{r}[v]$ of the transmitted symbol $\mathbf{d}[l]$ can be obtained by filtering $y[n]$ with a filter $g_R[n]$, circularly shifted in both the domains, as

$$\hat{r}[v] = \sum_{n=0}^{N-1} y[n] g_R[\lfloor n - \lfloor v \rfloor_M K \rfloor_N] \exp\left(j \frac{2\pi n}{N} (\lfloor v \rfloor_M - v)\right) \quad (4)$$

where $v = 0, 1, \dots, N-1$. Substituting (2) and (3) in (4) gives

$$\hat{r}[v] = \mathbf{d}[v] L(v, v) + \sum_{l=0, l \neq v}^{N-1} \mathbf{d}[l] L(v, l) + \sum_{n=0}^{N-1} z[n] \xi(v, n) \quad (5)$$

where the ICI coefficients $L(v, l)$ and $\xi(v, n)$ be denoted as

$$\begin{aligned} L(v, l) &= \sum_{n=0}^{N-1} g_R[\lfloor n - \lfloor v \rfloor_M K \rfloor_N] g_T[\lfloor n - \lfloor l \rfloor_M K \rfloor_N] \\ &\quad \times \exp\left(j \frac{2\pi n}{N} (M\epsilon + l - v + \lfloor v \rfloor_M - \lfloor l \rfloor_M)\right) \\ \xi(v, n) &= g_R[\lfloor n - \lfloor v \rfloor_M K \rfloor_N] \exp\left(j \frac{2\pi n}{N} (M\epsilon - v + \lfloor v \rfloor_M)\right) \end{aligned} \quad (6)$$

Here, ϵ represents the CFO normalised with K .

A. Offset QAM-GFDM

1) *TS-OQAM-GFDM*: Self interference free transmission is achieved by transmitting $d_{k,m}^{(i)}$ and $d_{k,m}^{(q)}$ as two real valued signals using half Nyquist prototype filters with an offset of $K/2$ samples from each other and a phase rotation of 90 degrees among adjacent subcarriers and subsymbols given as

$$\begin{aligned} g_{k,m}^{(i)}[n] &= \exp\left(j k \frac{\pi}{2}\right) \quad g_{k,m}^{(q)}[n] \\ g_{k,m}^{(q)}[n] &= \exp\left(j(k+1) \frac{\pi}{2}\right) g_{k,m+\frac{1}{2}}^{(i)}[n] \end{aligned} \quad (7)$$

which leads to

$$x[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{k,m}^{(i)} g_{k,m}^{(i)}[n] + \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{k,m}^{(q)} g_{k,m}^{(q)}[n] \quad (8)$$

Equation (8) can also be represented in a compact form as

$$x[n] = \sum_{l=0}^{N-1} \mathbf{d}^{(i)}[l] T(l, 0, 0) + \sum_{l=0}^{N-1} \mathbf{d}^{(q)}[l] T\left(l, \frac{1}{2}, 1\right) \quad (9)$$

where $\mathbf{d}^{(i)} = \Re\{\mathbf{d}\}$, $\mathbf{d}^{(q)} = \Im\{\mathbf{d}\}$ and

$$\begin{aligned} T\left(l, s_1, s_2\right) &= g_T[\lfloor n - \lfloor l + s_1 \rfloor_M K \rfloor_N] \\ &\quad \times \exp\left(j 2\pi \frac{n}{N} (l - \lfloor l \rfloor_M) + j \frac{\pi}{2} (s_2 + \lfloor l \rfloor_M - l)\right) \end{aligned} \quad (10)$$

After passing the TS-OQAM-GFDM signal through an AWGN channel, the transmitted symbol can be properly recovered by matched filtering with a filter which is circularly shifted in both the domains as $\hat{r}[v] = \Re\{\hat{r}_1[v]\} + j\Re\{\hat{r}_2[v]\}$ where

$$\hat{r}_1[v] = \sum_{n=0}^{N-1} y[n] T\left(v, 0, 0\right)^*; \quad \hat{r}_2[v] = \sum_{n=0}^{N-1} y[n] T\left(v, \frac{1}{2}, 1\right)^* \quad (11)$$

where $*$ denotes conjugate. Substituting (3) in (11) gives

$$\begin{aligned} \hat{r}_i[v] &= \mathbf{d}^{(i)}[v] L_{2i-1}(v, v) + \sum_{l=0, l \neq v}^{N-1} \mathbf{d}^{(i)}[l] L_{2i-1}(v, l) \\ &\quad + \mathbf{d}^{(q)}[v] L_{2i}(v, v) + \sum_{l=0, l \neq v}^{N-1} \mathbf{d}^{(q)}[l] L_{2i}(v, l) \\ &\quad + \sum_{n=0}^{N-1} (z[n] \xi_i(v, n)); \quad i \in \{1, 2\} \end{aligned} \quad (12)$$

where $L_1(v, l) = C(0, 0, 0, 0)$, $L_2(v, l) = C(\frac{1}{2}, 1, 0, 0)$ and $L_3(v, l) = C(0, 0, \frac{1}{2}, 1)$, $L_4(v, l) = C(\frac{1}{2}, 1, \frac{1}{2}, 1)$ and $\xi_1(v, n) = T(v, 0, 0)^*$, $\xi_2(v, n) = T(v, \frac{1}{2}, 1)^*$ and $C(s_1, s_2, s_3, s_4) = \sum_{n=0}^{N-1} T(l, s_1, s_2) T(v, s_3, s_4)$

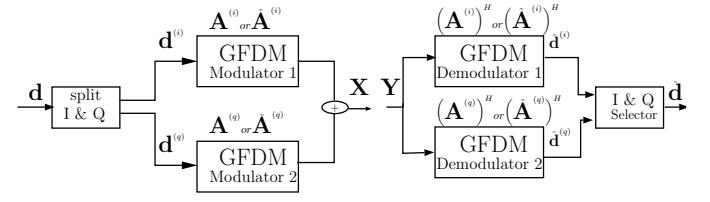


Fig. 1. Block diagram of TS/FS OQAM-GFDM

2) *FS-OQAM-GFDM*: The block diagram of TS/FS OQAM-GFDM is shown in Fig. 1. The transmitted waveform (8) can be written in a compact way as $\mathbf{x} = \mathbf{A}^{(i)} \mathbf{d}^{(i)} + \mathbf{A}^{(q)} \mathbf{d}^{(q)}$. The columns of matrix $\mathbf{A}^{(i)}$ and $\mathbf{A}^{(q)}$ carry $g_{k,m}^{(i)}$ and $g_{k,m}^{(q)}$. By considering the unitary DFT transform matrix $[\mathbf{W}_N]_{i,l} = 1/\sqrt{N} \exp(-j2\pi il/N)$, 2 precoding matrices are constructed from $\mathbf{A}^{(i)}$ and $\mathbf{A}^{(q)}$ as $\tilde{\mathbf{A}}^{(i)} = \mathbf{W}_N^H \mathbf{A}^{(i)}$; $\tilde{\mathbf{A}}^{(q)} = \mathbf{W}_N^H \mathbf{A}^{(q)}$. Using the above 2 matrices, the scheme was proposed as $\mathbf{x} = \tilde{\mathbf{A}}^{(i)} \mathbf{d}^{(i)} + \tilde{\mathbf{A}}^{(q)} \mathbf{d}^{(q)}$ in [6] which is equivalent to applying the modulation in frequency domain and then transforming to the time domain as

$$\hat{\mathbf{r}} = \Re\left\{\left(\tilde{\mathbf{A}}^{(i)}\right)^H \mathbf{y}\right\} + j\Re\left\{\left(\tilde{\mathbf{A}}^{(q)}\right)^H \mathbf{y}\right\} \quad (13)$$

where $\mathbf{y} = \mathbf{x} + \mathbf{z}$. The columns of $\tilde{\mathbf{A}}^{(i)}$ are the IDFTs of columns of $\mathbf{A}^{(i)}$ and the entries of $\tilde{g}_{k,m+c}$ are represented as

$$\tilde{g}_{k,m+c}[n] = G\left[\lfloor n + kM \rfloor_N\right] \exp\left(j 2\pi \frac{(m+c)n}{M}\right) \quad (14)$$

where $c \in \{0, 0.5\}$ and $G[n]$ is the frequency domain representation of $g_{k,m}[n]$. FS-OQAM-GFDM waveform is

$$x[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{k,m}^{(i)} \tilde{g}_{k,m}^{(i)}[n] + \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} d_{k,m}^{(q)} \tilde{g}_{k,m}^{(q)}[n] \quad (15)$$

which can also be represented in a compact way as

$$x[n] = \sum_{l=0}^{N-1} \mathbf{d}^{(i)}[l] F\left(l, 0, 0\right) + \sum_{l=0}^{N-1} \mathbf{d}^{(q)}[l] F\left(l, \frac{1}{2}, 1\right) \quad (16)$$

where

$$\begin{aligned} F\left(l, s_1, s_2\right) &= G\left[\lfloor n + FLR\left(\frac{l}{M}\right) M \rfloor_N\right] \exp\left(j 2\pi \frac{n}{M}\right. \\ &\quad \times \left.\left(\lfloor l + s_1 \rfloor_M + j \frac{\pi}{2} (s_2 + \lfloor l \rfloor_M - l)\right)\right) \end{aligned} \quad (17)$$

$F\text{LR}(\cdot)$ represents the floor operation. By replacing $T(\cdot)$ with $F(\cdot)$, TS-OQAM-GFDM converts to FS-OQAM-GFDM and the symbol error rate analysis is same for both the schemes.

III. SER ANALYSIS OF GFDM UNDER AWGN CHANNEL

A. BPSK Modulation

By following in a similar way as derived in [3][4], we obtain the probability of error expression for BPSK GFDM as

$$P_b(\xi) = \frac{1}{N2^{N-1}} \sum_{v=0}^{N-1} \sum_{l=0}^{2^N-2} Q\left(\sqrt{\frac{2\gamma}{E_\xi}} \alpha_{v,l}\right) + Q\left(\sqrt{\frac{2\gamma}{E_\xi}} \beta_{v,l}\right) \quad (18)$$

where $\gamma = \frac{E_b}{N_0}$ is the average energy per bit to noise spectral density ratio and $Q(\cdot)$ is the Gaussian Q function. The noise enhancement factor (NEF) is denoted as $E_\xi = \sum_{n=0}^{N-1} \xi(v, n)^2$ and $\alpha_{v,l} = \Re\{L(v, v) + \phi_v u_l\}$, $\beta_{v,l} = \Re\{L(v, v) - \phi_v u_l\}$, $\phi_v = (L(v, 0), L(v, 1), \dots, L(v, N-1)) \forall l \neq v$ and u_l is the l th column of a matrix \mathbf{U} with dimension $N-1 \times 2^{N-2}$ which is the binary representation of the number $2^{N-1} - l$ obtained by replacing zeros with negative 1s.

B. QPSK Modulation

1) *Normal QPSK*: The SER expression of normal QPSK GFDM is found out to be

$$P_s(\xi) = 1 - \frac{1}{N \times 2^{2(N-1)}} \sum_{v=0}^{N-1} \sum_{k=1}^{2^N-2} \sum_{n=1}^{2^N-2} \sum_{m=1}^4 Q(-a)Q(-b) \quad (19)$$

where $a = \sqrt{2\gamma/E_\xi} \Upsilon[1, m]$, $b = \sqrt{2\gamma/E_\xi} \Upsilon[2, m]$ and $\gamma = E_b/N_0$. Υ is a matrix of dimension 2×4 defined as $\Upsilon = [\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4]$ where $\mathbf{W}_1 = \mathbf{D} + \mathbf{W}_A + \mathbf{W}_B$, $\mathbf{W}_2 = \mathbf{D} - \mathbf{W}_A - \mathbf{W}_B$, $\mathbf{W}_3 = \mathbf{D} + \mathbf{W}_A - \mathbf{W}_B$, $\mathbf{W}_4 = \mathbf{D} - \mathbf{W}_A + \mathbf{W}_B$, $\mathbf{W}_A = \mathbf{L}_A \mathbf{e}_k$, $\mathbf{W}_B = \mathbf{L}_B \mathbf{e}_n$. $\mathbf{L}_A = [\mathbf{L}_A^0 \ \mathbf{L}_A^1 \ \dots \ \mathbf{L}_A^{N-1}]$, $\mathbf{L}_B = [\mathbf{L}_B^0 \ \mathbf{L}_B^1 \ \dots \ \mathbf{L}_B^{N-1}] \forall l \neq v$ and \mathbf{e}_k and \mathbf{e}_n are columns from a matrix of dimension $N-1 \times 2^{N-2}$. The k^{th} column is the binary representation of number $2^{N-1} - k$ where zeros are replaced with negative 1s. $\mathbf{L}_A^l = [\Re\{L(v, l)\} \ \Im\{L(v, l)\}]^T$, $\mathbf{D} = [\Re\{L(v, v)\} - \Im\{L(v, v)\} \ \Re\{L(v, v)\} + \Im\{L(v, v)\}]^T$ and $\mathbf{L}_B^l = [\Im\{L(v, l)\} - \Re\{L(v, l)\}]^T$. $[\cdot]^T$ denotes the transpose operation.

2) *Offset QPSK*: The SER expression for offset QPSK GFDM is found out to be similar to SER expression of normal QPSK GFDM except for change in $a = \sqrt{2\gamma/E_\xi} \Upsilon[1, m]$, $b = \sqrt{2\gamma/E_\xi} \Upsilon[2, m]$, $\mathbf{L}_A^l = [\Re\{L_1(v, l)\} \ \Re\{L_3(v, l)\}]^T$, $\mathbf{D} = [\Re\{L_1(v, v) + L_2(v, v)\} \ \Re\{L_3(v, v) + L_4(v, v)\}]^T$ and $\mathbf{L}_B^l = [\Re\{L_2(v, l)\} \ \Re\{L_4(v, l)\}]^T$. The NEF is denoted as $E_{\xi_1} = \sum_{n=0}^{N-1} \xi_1(v, n)^2$, $E_{\xi_2} = \sum_{n=0}^{N-1} \xi_2(v, n)^2$.

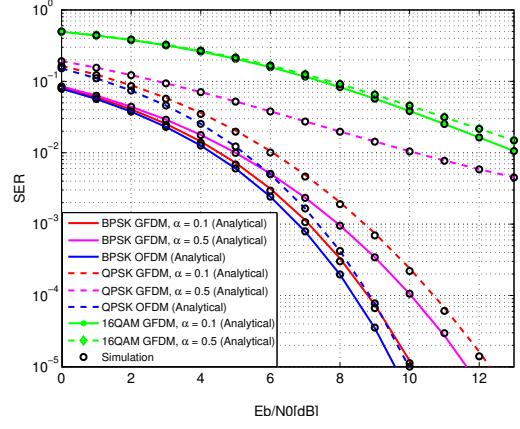


Fig. 2. SER performance of normal QAM GFDM for $K=3$, $M=3$, $\epsilon = 0.01$ for BPSK & QPSK, $\epsilon = 0.001$ for 16QAM

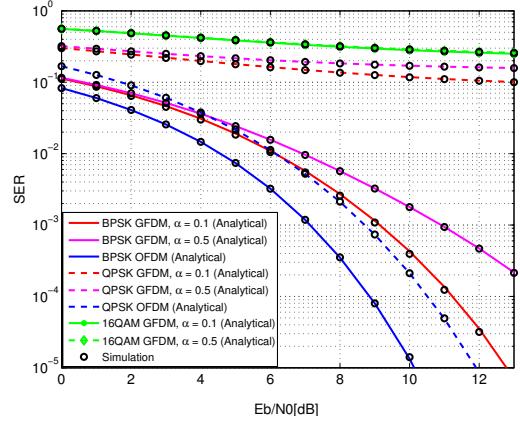


Fig. 3. SER performance of normal QAM GFDM for $K=3$, $M=3$, $\epsilon = 0.05$ for BPSK & QPSK, $\epsilon = 0.005$ for 16QAM

C. 16-QAM Modulation

1) *Normal 16 QAM*: The SER expression for normal 16 QAM GFDM is found out to be

$$P_s(\xi) = 1 - \frac{1}{4N \times 2^{4(N-1)}} \sum_{v=0}^{N-1} \sum_{k=1}^{2^N-2} \sum_{n=1}^{2^N-2} \sum_{p=1}^{2^N-2} \sum_{q=1}^{2^N-2} \sum_{m=1}^{16} Q\left(\bar{\gamma}_1 \mathbf{B}_1^1(0)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^1(0)\right) - Q\left(\bar{\gamma}_1 \mathbf{B}_1^1(0)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^1(2)\right) - Q\left(\bar{\gamma}_1 \mathbf{B}_1^1(2)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^1(0)\right) + Q\left(\bar{\gamma}_1 \mathbf{B}_1^1(2)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^1(2)\right) + Q\left(\bar{\gamma}_1 \mathbf{B}_1^2(0)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^2(2)\right) - Q\left(\bar{\gamma}_1 \mathbf{B}_1^2(2)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^2(2)\right) + Q\left(\bar{\gamma}_1 \mathbf{B}_1^3(2)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^3(0)\right) - Q\left(\bar{\gamma}_1 \mathbf{B}_1^3(2)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^3(2)\right) + Q\left(\bar{\gamma}_1 \mathbf{B}_1^4(2)\right) Q\left(\bar{\gamma}_2 \mathbf{B}_2^4(2)\right) \quad (20)$$

where $\bar{\gamma}_1 = \bar{\gamma}_2 = -\sqrt{4\gamma/5E_\xi}$ and $\mathbf{B}_1^i(a) = \mathbf{A}_{knpqv}^i[1, m] - a$, $\mathbf{B}_2^i(a) = \mathbf{A}_{knpqv}^i[2, m] - a$. $\mathbf{A}_{knpqv}^1 = (1\mathbf{L}_A^v - 1\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{A}_{knpqv}^2 = (1\mathbf{L}_A^v - 3\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{A}_{knpqv}^3 = (3\mathbf{L}_A^v - 1\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{A}_{knpqv}^4 = (3\mathbf{L}_A^v - 3\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{C} = [\mathbf{L}_A \mathbf{e}_k \ 2\mathbf{L}_A \mathbf{e}_n \ \mathbf{L}_B \mathbf{e}_p \ 2\mathbf{L}_B \mathbf{e}_q] \forall l \neq v$. The column vectors

$\mathbf{e}_k, \mathbf{e}_n, \mathbf{e}_p, \mathbf{e}_q$ are taken from matrix E_{M-1} of order $M - 1 \times 2^{M-2}$. The n th column of E_M is the binary representation of $2^M - n$ where binary zeros are changed with negative ones. \mathbf{d}_r is the r th column of matrix of dimension 4×2^4 which represents $(16 - r)$ in binary format where binary zeros are replaced with negative 1s.

2) *Offset 16 QAM*: Following in a similar way as derived for normal 16 QAM-GFDM, we obtain a SER expression similar to it except for change in $\bar{\gamma}_1 = -\sqrt{4\gamma/5E_{\xi_1}}$, $\bar{\gamma}_2 = -\sqrt{4\gamma/5E_{\xi_2}}$, $\mathbf{L}_A^l = [\Re\{L_1(v, l)\} \ \Re\{L_3(v, l)\}]^T$, $\mathbf{D} = [\Re\{L_1(v, v) + L_2(v, v)\} \ \Re\{L_3(v, v) + L_4(v, v)\}]^T$, $\mathbf{L}_B^l = [\Re\{L_2(v, l)\} \ \Re\{L_4(v, l)\}]^T$, $\mathbf{A}_{knpqv}^1 = (1\mathbf{L}_A^v + 1\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{A}_{knpqv}^2 = (1\mathbf{L}_A^v + 3\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{A}_{knpqv}^3 = (3\mathbf{L}_A^v + 1\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$, $\mathbf{A}_{knpqv}^4 = (3\mathbf{L}_A^v + 3\mathbf{L}_B^v + \mathbf{C}\mathbf{d}_m)$

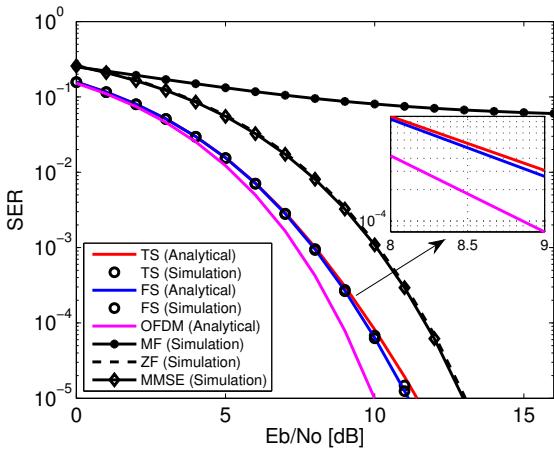


Fig. 4. SER performance of offset QPSK-GFDM for $\epsilon = 0.01$, $K=4$, $M=3$ and for OFDM ($K=12$). TS means TS-OQAM-GFDM and FS means FS-OQAM-GFDM

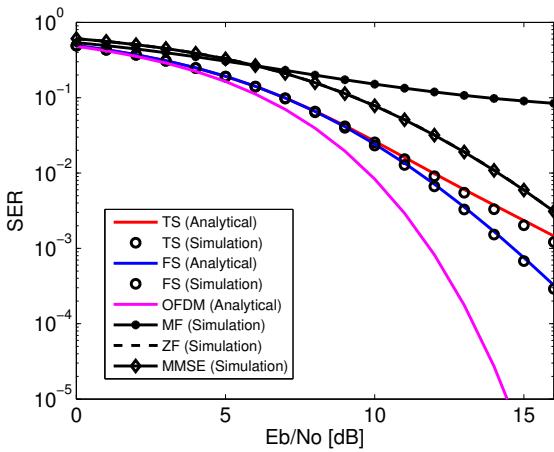


Fig. 5. SER performance of offset 16-QAM GFDM for $\epsilon = 0.01$, $K=2$, $M=3$ and for OFDM ($K=6$). TS means TS-OQAM-GFDM and FS means FS-OQAM-GFDM

IV. SER ANALYSIS OF GFDM UNDER TWDP

The approximate probability density (pdf) function of TWDP fading channel is given in [7] as

$$f_R(r) = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2} - \bar{K}\right) \sum_{i=1}^M \frac{a_i}{2} \times \left\{ \exp(+\alpha_i \bar{K}) I_0\left(\frac{r}{\sigma} \sqrt{2\bar{K}(1-\alpha_i)}\right) + \exp(-\alpha_i \bar{K}) I_0\left(\frac{r}{\sigma} \sqrt{2\bar{K}(1+\alpha_i)}\right) \right\} \quad (21)$$

where random variable (RV) $r \in \mathbb{R}$ is the fading envelope which is TWDP distributed. $I_0(\cdot)$ represents zeroth-order modified bessel function and $2\sigma^2$ represents the average power of the diffused waves. $\bar{K} = (V_1^2 + V_2^2)/2\sigma^2$ represents the ratio of total specular power to the diffused power. $\Delta = 2V_1V_2/(V_1^2 + V_2^2)$ indicates the relative strength of both specular components and $\alpha_i = \Delta \cos(\pi(i-1)/(2M-1))$. M is the order of the TWDP pdf and as a general rule of thumb, $M \geq \frac{1}{2}\bar{K}\Delta$ should be used so that (21) does not significantly differ from the actual TWDP pdf. $a_i = 2(-1)^i \int_0^{2M-1} \prod_{k=1, k \neq i}^{2M} (u-k+1) du / ((2M-1)(2M-i)!(i-1)!)$. When $\bar{K} = 0$, TWDP pdf converts to Rayleigh pdf where only the diffused waves exist and when $\bar{K} \neq 0$ and $\Delta = 0$, it converts to Rician pdf. When $V_1 = V_2$, $\Delta = 1$, the channel fades poorer than Rayleigh channel for large \bar{K} .

A. BPSK-GFDM

In the presence of fading, the conditional bit error probability equation can be written using (18) as

$$P_b(\xi|r) = \frac{1}{N2^{N-1}} \sum_{v=0}^{N-1} \sum_{l=0}^{2^{N-2}} Q\left(\sqrt{\frac{2\gamma}{E_\xi}} \alpha_{v,l} r\right) Q\left(\sqrt{\frac{2\gamma}{E_\xi}} \beta_{v,l} r\right) \quad (22)$$

The bit error probability is defined as $P_b(\xi) = \int_0^\infty P_b(\xi|r) f_R(r) dr$. By using the identities (24), (25) and following in a similar way as in [2], the BER expression is found out to be

$$P_b(\xi) = \frac{1}{N2^N} \sum_{i=1}^M a_i \sum_{n=0}^\infty \sum_{v=0}^{N-1} \sum_{l=0}^{2^{N-2}} \left\{ \exp(-\bar{K}(1-\alpha_i)) \times \left(S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1-\alpha_i); \frac{2\sigma^2\gamma\alpha_{v,l}^2}{E_\xi}\right) + S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1-\alpha_i); \frac{2\sigma^2\gamma\beta_{v,l}^2}{E_\xi}\right) \right) \right\} \\ + \left\{ \exp(-\bar{K}(1+\alpha_i)) \times \left(S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1+\alpha_i); \frac{2\sigma^2\gamma\alpha_{v,l}^2}{E_\xi}\right) + S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1+\alpha_i); \frac{2\sigma^2\gamma\beta_{v,l}^2}{E_\xi}\right) \right) \right\} \quad (23)$$

$$\begin{aligned}
P_s(\xi) = & \frac{1}{N2^{2N}} \sum_{v=0}^{N-1} \sum_{k=1}^{2^{N-2}} \sum_{p=1}^{2^{N-2}} \sum_{m=1}^4 \sum_{i=1}^M a_i \\
& \times \left\{ \exp(-\bar{K}(1-\alpha_i)) \sum_{n=0}^{\infty} \left\{ 2S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1-\alpha_i); a^2\sigma^2\right) + 2S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1-\alpha_i); b^2\sigma^2\right) \right. \right. \\
& \quad \left. \left. - S_{n+1}\left(\theta_1; \bar{K}(1-\alpha_i); a^2\sigma^2\right) - S_{n+1}\left(\theta_2; \bar{K}(1-\alpha_i); b^2\sigma^2\right) \right\} \right. \\
& + \exp(-\bar{K}(1+\alpha_i)) \sum_{n=0}^{\infty} \left\{ 2S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1+\alpha_i); a^2\sigma^2\right) + 2S_{n+1}\left(\frac{\pi}{2}; \bar{K}(1+\alpha_i); b^2\sigma^2\right) \right. \\
& \quad \left. \left. - S_{n+1}\left(\theta_1; \bar{K}(1+\alpha_i); a^2\sigma^2\right) - S_{n+1}\left(\theta_2; \bar{K}(1+\alpha_i); b^2\sigma^2\right) \right\} \right\} \\
(26)
\end{aligned}$$

where $S_{n+1}(\theta; b; c)$ is shown to converge faster than the normal exponential series and

$$\sum_{n=0}^{\infty} S_{n+1}(\theta; b; c) = \sum_{n=0}^{\infty} \frac{b^n}{n!} G_{n+1}(\theta; c) \quad (24)$$

$$\begin{aligned}
G_{n+1}(\theta; c) &= \frac{1}{\pi} \int_{\phi=0}^{\theta} \left(\frac{\sin^2 \phi}{\sin^2 \phi + c} \right)^{n+1} d\phi \\
&= \frac{\theta}{\pi} - \frac{\beta}{\pi} \left[\left(\frac{\pi}{2} + \tan^{-1}(\alpha) \right) \sum_{k=0}^n \binom{2k}{k} \right. \\
&\quad \times (4+4c)^{-k} + \sin(\tan^{-1}(\alpha)) \sum_{k=1}^n \sum_{i=1}^k \frac{T_{ik}}{(1+c)^k} \\
&\quad \left. \times \cos(\tan^{-1}(\alpha))^{2(k-i)+1} \right] \\
(25)
\end{aligned}$$

where $\beta = \sqrt{c/(1+c)} \operatorname{sgn}(\theta)$, $\alpha = -\beta \cot(\theta)$ and $T_{ik} = \binom{2k}{k} / \left(\binom{2(k-i)}{k-i} 4^i 2(k-i) + 1 \right)$.

B. QPSK-GFDM

By following in a similar way as derived for BPSK-TWDP fading and using the identity (27), the symbol error probability for QPSK-TWDP fading is given as (26) where

$$\begin{aligned}
\theta_1 &= \frac{\pi}{2} - \tan^{-1}\left(\frac{b}{a}\right), \theta_2 = \tan^{-1}\left(\frac{b}{a}\right), \\
Q(ar)Q(br) &= \frac{1}{2\pi} \int_0^{\frac{\pi}{2}-\tan^{-1}(\frac{b}{a})} \exp\left(-\frac{a^2 r^2}{2 \sin^2 \theta}\right) d\theta \\
&+ \frac{1}{2\pi} \int_0^{\tan^{-1}(\frac{b}{a})} \exp\left(-\frac{b^2 r^2}{2 \sin^2 \theta}\right) d\theta
(27)
\end{aligned}$$

V. SIMULATION RESULTS AND ANALYSIS

The numerical and simulation analysis is done for normal BPSK, QPSK and 16 QAM GFDM with root raised cosine (RRC) filter as the prototype filter for different values of roll off

factor (α) and for varying frequency offset (ϵ). The simulation parameters used in Fig. 2 and Fig. 3 are $K = 3$, $M = 3$, $\epsilon = 0.01$ for BPSK, QPSK GFDM, $\epsilon = 0.001$ for 16-QAM GFDM, and $K = 9$ for OFDM. There is a strong influence of α on the SER performance as seen in Fig. 2 and Fig. 3. SER increases with increase in α because self interference increases in normal QAM GFDM. As the modulation order μ increases, there is a strong deviation of SER graph from OFDM graph even for low values of α . It is observed that for low values of α , SER of GFDM and OFDM are nearly equal for low values of ϵ . As μ increases, there is a noticeable difference between the two. It can be seen that our results match to the SER expressions of OFDM in [3] for $M = 1$, $N = K$ and $g_T = g_R = 1$.

The self interference problem in normal QAM GFDM is solved by using either TS-OQAM or FS-OQAM approach. The analytical and simulation analysis is done for both TS-OQAM-GFDM and FS-OQAM-GFDM, in general for offset QPSK & offset 16 QAM. In normal QAM GFDM, there is a strong influence of α on the SER performance, but in OQAM-GFDM it is negligible. The SER performance analysis is shown in Fig. 4 and Fig. 5 for $\alpha = 0.9$, $\epsilon = 0.01$, $K = 4$, $M = 3$ for offset QPSK GFDM and $K = 2$, $M = 3$ for offset 16-QAM GFDM and $K = [12, 6]$ for OFDM. It is observed that for higher values of α , the performance of FS-OQAM-GFDM is better than conventional MF, ZF and MMSE receivers. It is observed that there is a noticeable difference between SER performance of TS-OQAM-GFDM and FS-OQAM-GFDM for higher values of ϵ . From Fig. 4 and 5, it can be seen that FS-OQAM-GFDM slightly outperforms TS-OQAM-GFDM at high SNR for $\epsilon = 0.01$. In normal QAM GFDM there is a performance degradation when even number of subsymbols are transmitted, but in OQAM-GFDM it is not seen. It is observed that the computational complexity increases as N increases. So, the results are evaluated for smaller values of N to check the accuracy of the theoretical expression.

Fig. 6 shows the BER performance of BPSK GFDM under TWDP fading for $\alpha = 0.1$, $\epsilon = 0.01$, $\Delta = 1.0$ with varying

\bar{K} . It is observed that for $\bar{K} > 6$ dB, the BER performance of TWDP fading is poorer than Rayleigh fading. It can be observed from Fig. 7 that for $\bar{K} > 10$ dB and $\Delta = 1.0$, the BER performance is poorer than Rayleigh fading. Since the performance of OQAM GFDM is better than normal QAM GFDM under AWGN channel, the SER performance analysis is done for OQPSK GFDM under TWDP fading for different values of Δ and \bar{K} for $\alpha = 0.9$ and $\epsilon = 0.01$. Fig. 8 shows that for $\Delta = 0$ and $\bar{K} = 10$, the SER performance is good and for $\Delta = 1$ and $\bar{K} = 10$, the SER performance is poor.

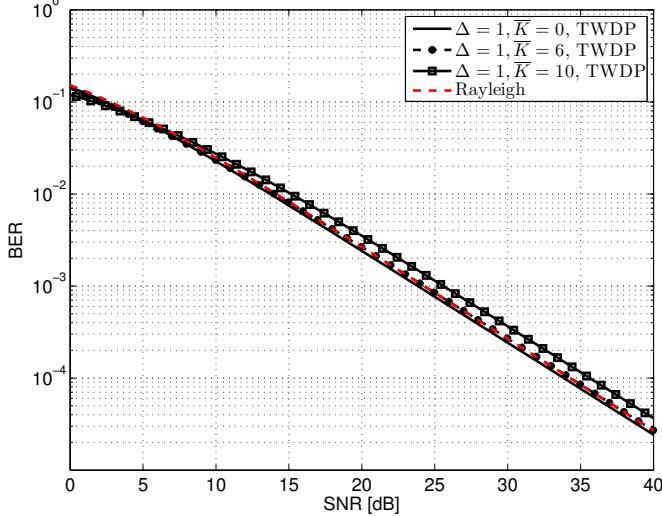


Fig. 6. BPSK-GFDM receiver performance for $\alpha = 0.1$, $\epsilon = 0.01$, $\Delta = 1.0$, varying \bar{K}

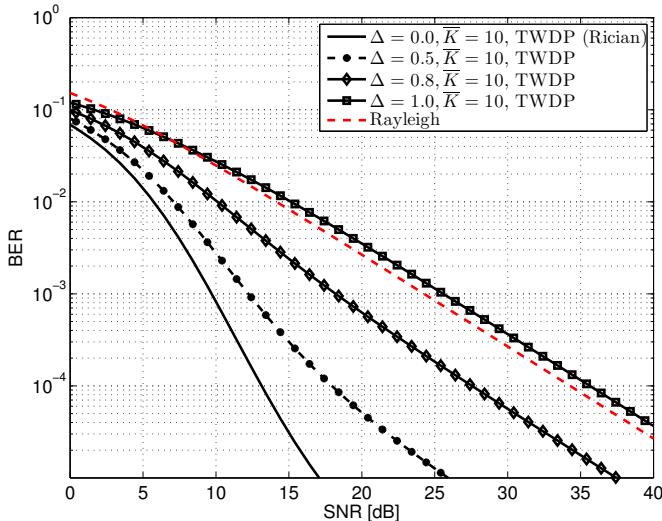


Fig. 7. BPSK-GFDM receiver performance for $\alpha = 0.1$, $\epsilon = 0.01$, $\bar{K} = 10$ dB, varying Δ

VI. CONCLUSION

Closed form SER expressions for normal QAM GFDM and offset QAM GFDM system with CFO have been derived under AWGN channel and TWDP fading channel. For normal QAM GFDM, under AWGN channel it is observed that

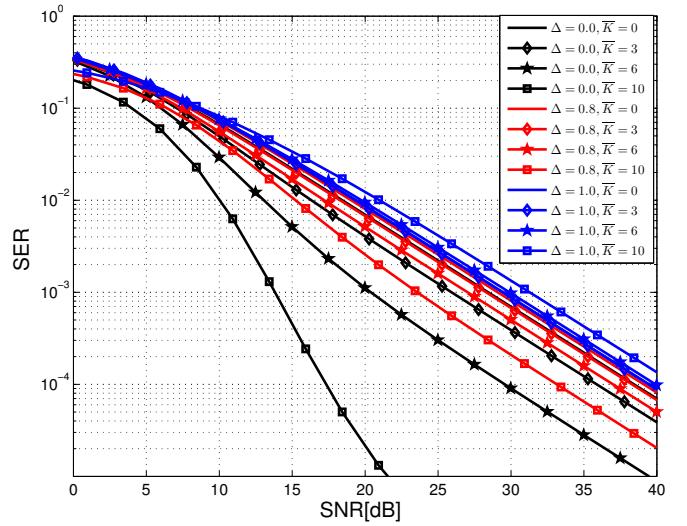


Fig. 8. OQPSK-GFDM receiver performance for $\alpha = 0.9$, $\epsilon = 0.01$, varying \bar{K} , varying Δ

SER performance of the matched filter receiver is close to ideal waveform at low SNR for low values of α , ϵ and μ . The SER performance of offset QAM GFDM is better than normal QAM GFDM because orthogonality is preserved and FS-OQAM-GFDM slightly outperforms TS-OQAM-GFDM at high SNR for higher values of CFO. FS-OQAM-GFDM outperforms conventional MF, ZF and MMSE receivers for higher values of α . Two specular components which are equal in strength and opposite in phase can result in a performance poorer than Rayleigh fading. Since the futuristic waveforms have to be less complex and have to perform better for higher modulation order for higher data rates, offset QAM approach is the preferred choice.

REFERENCES

- [1] N. Michailow et al., "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks" in *IEEE Transactions on Communications*, vol. **62**, no. 9, pp. 3045-3061, Sept. 2014
- [2] B. Farhang-Boroujeny, A. Farhang, A. RezazadehReyhani, A. Aminjaveri and D. Qu, "A comparison of linear FBMC and circularly shaped waveforms," in 2016 IEEE/ACES International Conference on Wireless Information Technology and Systems (ICWITS) and Applied Computational Electromagnetics (ACES), Honolulu, HI, 2016, pp. 1-2.
- [3] P. Dharmaawansa, N. Rajatheva and H. Minn, "An exact error probability analysis of OFDM systems with frequency offset" in *IEEE Transactions on Communications*, vol. **57**, no. 1, pp. 26-31, January 2009.
- [4] D. Gaspar, L. Mendes and T. Pimenta, "GFDM BER Under Synchronization Errors" in *IEEE Communications Letters*, vol. **21**, no. 8, pp. 1743-1746, Aug. 2017.
- [5] N. Michailow, S. Krone, M. Lentmaier and G. Fettweis, "Bit Error Rate Performance of Generalized Frequency Division Multiplexing," 2012 IEEE Vehicular Technology Conference (VTC Fall), Quebec City, QC, 2012, pp. 1-5.
- [6] Ivan Gaspar et al., "Frequency-Shift Offset-QAM for GFDM" in *IEEE Communications Letters*, vol. **19**, no. 8, pp. 1454-1457, Aug. 2015.
- [7] G. D. Durgin, T. S. Rappaport and D. A. de Wolf, "New analytical models and probability density functions for fading in wireless communications," in *IEEE Transactions on Communications*, vol. **50**, no. 6, pp. 1005-1015, Jun 2002.
- [8] M. K. Simon and M. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels," in *Proceedings of the IEEE*, vol. **86**, no. 9, pp. 1860-1877, Sep 1998.

Performance Analysis of Wireless Powered Decode-and-Forward Relay System

Pawan Kumar

Department of Electronics and Electrical Engineering
Indian Institute of Technology - Guwahati
Guwahati 781039, India
E-mail: kpawan@iitg.ac.in

Kalpana Dhaka

Department of Electronics and Electrical Engineering
Indian Institute of Technology - Guwahati
Guwahati 781039, India
E-mail: kalpana.dhaka@iitg.ac.in

Abstract—We consider a two-hop decode-and-forward relay system with wireless powered source node. Source and destination are in the coverage range of the relay and direct link connecting them is assumed to be blocked. The signal transmitted by the relay node is used to harvest energy at the source and forward data to the destination in even time slot. In the following odd slot the energy harvested at source is used to communicate data to relay node. The cycle continues and the source data is communicated to the blocked destination node. The average symbol error rate (SER) of the system is analyzed for Rayleigh faded links when data is i) M -ary phase-shift keying modulated with coherent detection and ii) orthogonal M -ary frequency-shift keying modulated with non-coherent detection. The high signal-to-noise ratio approximations of the average SER are also obtained to investigate the effects of modulation order and relay placement on the system's performance.

I. INTRODUCTION

Recently radio frequency (RF)-based energy harvesting (EH) is widely explored as a potential solution to wirelessly power energy-constrained nodes [1]–[4]. The major literature on wireless EH can be grouped into i) improving circuit design for EH [1]; ii) novel techniques for RF-based EH systems [2]–[6], and iii) performance analysis of RF-based EH systems [4], [6]–[9]. In a communication system, its capability and reliability are measured in terms of achievable throughput, outage, error-rate, etc. The commonly used approaches for performance analysis are broadly categorized into information theoretic and communication theoretic approach [8]. In [7], [9], the performance of a point-to-point communication system is analyzed. The performance of relay-assisted systems is investigated in [4], [6], [8]. Relay-assisted communication provides benefits in terms of improved reliability, increased coverage, etc. Amplify-and-forward (AF) and decode-and-forward (DF) are two main relaying protocols [10].

In the literature on wireless powered (WP) relay systems, relay node(s) are mainly considered to be energy-constrained and powered by source node(s), interferes, or dedicated power transmitter(s) [4], [6], [8]. However, relay node(s) can also be considered as an energy supplier while broadcasting when it is in the range of source and destination node(s). Other scenarios considered in the literature include systems where the destination node is energy-constrained and wirelessly powered by relay node(s), source node(s), and/or interferers to extend the

communication lifetime [11], [12]. Relatively less literature is available on systems where source node is energy-constrained. The analysis for such systems can be vital in applications like wireless sensor networks, wireless body area networks, machine-to-machine communication, and the internet of things (IoT) entities, such as smart home, medical implant, etc., [2], [3]. The energy-constrained source node first needs to harvest energy before it starts transmitting data. In [13]–[16], energy-constrained source node(s) rely on RF signals transmitted by relay and/or destination node(s) for EH. Information theoretic approach is considered to investigate the performance of the system in terms of throughput. AF protocol is employed at the relay in [13], whereas DF protocol is considered in [14]–[16]. In [13]–[15], the effect of fading is either ignored or assumed constant. In a practical scenario, fading can have severe consequences on the performance and hence its effects cannot be ignored. Moreover, the analytical results obtained without considering fading can largely deviate from the actual results. Mixed Rayleigh and Rician fading environment are considered in [16] to investigate the system throughput.

In this paper, we analyze the average SER of a two-hop half-duplex DF relay system where the direct link is blocked and energy-constrained source node harvests energy using RF signal transmitted by the relay node. Links are considered to be modeled using Rayleigh fading. Source and destination are in the coverage range of the relay and therefore they receive RF signal broadcasted by the relay. Source utilizes the received RF signal for EH, while the symbol is detected at the destination. The energy harvested at source is used for data transmission in the following slot. The process is pursued in an iterative manner to transmit data from source to destination. The main contributions of this paper include i) derivation of the average symbol error rate (SER) expressions for M -ary phase-shift keying (M -PSK) and orthogonal M -ary frequency-shift keying (M -FSK) modulated data, ii) high signal-to-noise ratio (SNR) approximations of the average SERs to find the optimal relay location and diversity order (DO) of the system, and iii) examining the effects of modulation scheme/order, relay location, and fading parameter on the system performance.

II. SYSTEM MODEL

The system consists of an energy-constrained source node S , a relay node R , and a destination node D . We consider

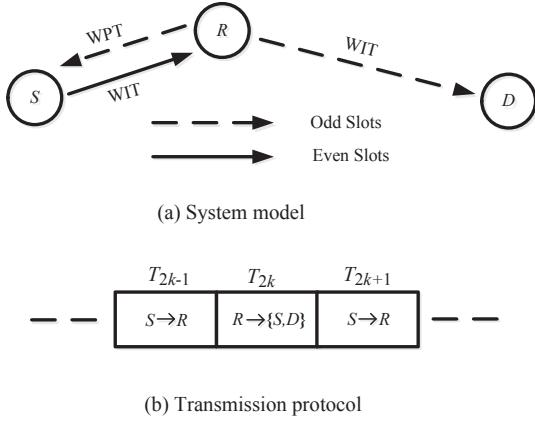


Fig. 1. The system model and the transmission protocol. Acronyms: WPT – wireless power transfer, WIT – wireless information transfer.

single antenna at each node and the links are Rayleigh faded. The direct link between nodes S and D is deeply faded and therefore it is ignored. The mode of communication is assumed to be half-duplex. Fixed DF protocol is employed at the relay node to process and forward the data at node R . Since nodes S and D are in the coverage range of node R ; the signal transmitted by node R is received at nodes S and D . Node D detects the transmitted data whereas at node S the signal is used for EH. The harvested energy is employed for data transfer in the next slot. The system model is illustrated in Fig. 1. The node R is assumed to have sufficient power supply for the end-to-end communication. We consider the harvest-use approach to process the incoming energy flow, that is, node S cannot store the harvested energy beyond the current time slot owing to hardware constraints [5]. In addition, we assume that power consumptions in operational circuits such as encoder, modulator, etc., is comparatively small and therefore can be neglected [17].

We consider a traditional two-hop DF relay system in which data is communicated from node S to node D in two time slots. In addition, we consider that the data at node S is communicated using the energy harvested at the node in the previous slot. In odd time slots, node S transmits data to node R using the energy harvested in the previous slot and in even time slots node R sends data which is received by nodes S and D . That is, in the odd time slot $(2k-1)$ (k is a positive integer), node S transmits data to node R using energy harvested in the previous time slot. Further, in the even time slot $2k$, node R broadcasts RF signal which on reception at node S and node D is used for EH and detection, respectively. The process is repeated in the following odd and even slots. We assume that all slots are of equal duration T_s , where T_s is symbol duration. For notational convenience, we refer odd and even time slots as $T_O = [2(k-1)T_s, (2k-1)T_s]$ and $T_E = [(2k-1)T_s, 2kT_s]$, respectively for $k = 1, 2, \dots$.

We consider two signaling schemes M -PSK and orthogonal M -FSK for data transmission. At receiving nodes, coherent detection is performed for orthogonal M -PSK and non-coherent detection for M -FSK. Each symbol has the support of

$[0, T_s]$. The symbols are transmitted with equal *a priori* probability. The constellation \mathcal{S} is given by $\mathcal{S} = \{S_1, \dots, S_M\}$, where S_1, \dots, S_M are the constellation points corresponding to M symbols $1, \dots, M$, respectively. In the case of M -PSK, the constellation is given by

$$S_m = \exp(j2\pi(m-1)/M), \quad m = 1, \dots, M, \quad (1)$$

where $j = \sqrt{-1}$. The orthogonal M -FSK constellation can be represented as

$$S_m = \exp(j2\pi(\Delta f_m)t), \quad m = 1, \dots, M, \quad (2)$$

where Δf_m denotes the shift of the m th frequency from the carrier such that

$$\frac{1}{T_s} \int_0^{T_s} S_m S_u^* dt = \begin{cases} 1, & m = u \\ 0, & m \neq u \end{cases}. \quad (3)$$

We assume links between nodes are independent, slow-flat Rayleigh faded.

The baseband equivalent of received RF signal at node R in slot T_O is given by

$$y_{SR}^O = \sqrt{P_S T_s (d_{SR})^{-\alpha_{SR}}} h_{SR} x + n_R^O, \quad (4)$$

where $x \in \mathcal{S}$ is the symbol transmitted by node S with power P_S , h_{SR} is the fading coefficient of SR link, $(d_{SR})^{-\alpha_{SR}}$ represents path loss of SR link, d_{SR} is distance in meters with path loss exponent α_{SR} , and n_R^O is additive white Gaussian noise (AWGN) at node R . Here, P_S depends on the energy harvested at node S in the previous even time slot. We assume that the energy harvested at node S in even slots is completely used to send data to node R in odd slots. Since node S relies on the energy harvested in the previous slot for data transmission, an arbitrary RF signal can be transmitted by the relay (say, in time slot T_0) to start the communication.

Further, in the subsequent slot T_E , the baseband equivalent of the signal communicated by node R to node j is

$$y_{Rj}^E = \sqrt{P_R T_s (d_{Rj})^{-\alpha_{Rj}}} h_{Rj} \hat{x} + n_j^E, \quad j = S, D \quad (5)$$

where \hat{x} is the estimate of symbol x transmitted in slot T_O , P_R is power transmitted by node R , h_{Rj} is the fading coefficient of Rj link, $(d_{Rj})^{\alpha_{Rj}}$ represents path loss of Rj link, d_{Rj} is distance in meters with path loss exponent α_{Rj} , and n_j^E is AWGN at node j . Path loss exponent ranges from 2 to 6 based on environmental conditions. The noise components n_R^O and n_j^E are independent of each other and the channel fading coefficients. They are zero-mean complex Gaussian with power spectral density (PSD) N_0 .

For a comparatively small noise component, the energy harvested at node S in slot T_E can be approximated as

$$E_S \approx \eta P_R T_s |h_{RS}|^2 (d_{RS})^{-\alpha_{RS}}, \quad (6)$$

where η is energy conversion efficiency. Since, $E_S = P_S T_s$, (6) can be re-written as

$$P_S \approx \eta P_R |h_{RS}|^2 (d_{RS})^{-\alpha_{RS}}. \quad (7)$$

In the case of M -PSK, the detected data \hat{x} resulting from coherent detection at node R in slot T_O is obtained from the decision rule

$$\hat{x} = \arg \left\{ \max_{s \in \mathcal{S}} \operatorname{Re} (s^* h_{SR}^* y_{SR}^O) \right\}, \quad (8)$$

where $\text{Re}(\cdot)$ represents the real component of the argument. At the destination, we get the detected symbol \hat{x} in slot T_E , given by

$$\hat{x} = \arg \left\{ \max_{s \in \mathcal{S}} \text{Re} (s^* h_{RD}^* y_{RD}^E) \right\}. \quad (9)$$

In case of orthogonal M -FSK, the detected data \hat{x} resulting from non-coherent detection at node R in slot T_O is obtained from the decision rule

$$\hat{x} = \arg \left\{ \max_{s \in \mathcal{S}} |s^* y_{SR}^O| \right\}, \quad (10)$$

and, finally, at the destination, we get the detected symbol \hat{x} in slot T_E , given by

$$\hat{x} = \arg \left\{ \max_{s \in \mathcal{S}} |s^* y_{RD}^E| \right\}. \quad (11)$$

The instantaneous received SNRs at nodes R and D can be expressed using (4), (5), and (7) as

$$\gamma_{SR} = \frac{\eta P_R T_s |h_{RS}|^2 |h_{SR}|^2}{d_{RS}^{\alpha_{RS}} d_{SR}^{\alpha_{SR}} N_0} \quad \text{and} \quad (12a)$$

$$\gamma_{RD} = \frac{P_R T_s |h_{RD}|^2}{d_{RD}^{\alpha_{RD}} N_0}, \quad (12b)$$

respectively. For Rayleigh fading, the power variates of the channel coefficients $|h_{ij}|^2$ are exponentially distributed with probability density function (PDF)

$$f_{|h_{ij}|^2}(h) = \frac{1}{\lambda_{ij}} \exp\left(-\frac{h}{\lambda_{ij}}\right). \quad (13)$$

The mean power of ij link is $E[|h_{ij}|^2] = \lambda_{ij}$, where $E[\cdot]$ represents expectation operator. Hence, the average received SNRs at nodes R and D are

$$\bar{\gamma}_{SR} = \frac{\eta \lambda_{RS} \lambda_{SR} P_R T_s}{d_{SR}^{(\alpha_{SR} + \alpha_{RS})} N_0}, \quad \text{and} \quad (14a)$$

$$\bar{\gamma}_{RD} = \frac{\lambda_{RD} P_R T_s}{(d_{RD})^{\alpha_{RD}} N_0}, \quad (14b)$$

respectively.

III. PERFORMANCE ANALYSIS

In this section, we present the average SER analysis for the system in Fig. 1 for i) M -PSK modulation scheme with coherent detection and ii) orthogonal M -FSK modulation scheme with non-coherent detection. These signaling schemes have symmetric constellations.

The end-to-end error probability for the system can be defined in terms of the probability of correct decision. The end-to-end probability of correct reception is a function of the possibilities when symbol S_p transmitted by node S is detected as S_q at node R , which when forwarded by R is detected as S_p at node D . Note that $S_p, S_q, S_r \in \mathcal{S}$. The symbol transmitted by node S can be correctly detected at node D , only if, $S_p = S_r$. Symbol S_p is transmitted by node S in slot T_O and detected as S_q at node R which is then transmitted by node R and detected as S_r at node D in slot T_E . Let $P_{p,q}(\gamma_\varphi)$ represents the conditional probability to send symbol S_p and receive S_q over link φ , where $S_p, S_q \in \mathcal{S}$ and $\varphi \in \{SR, RD\}$. $P_{p,q}(\gamma_\varphi)$ for $p \neq q$ corresponds to the conditional paired error probabilities, conditioned on the instantaneous SNR γ_φ [18], [19].

The conditional end-to-end probability of correct detection at node D can be stated as [18]

$$P_c(\gamma_{SR}, \gamma_{RD}) = \frac{1}{M} \sum_{\ell=1}^M \sum_{m=1}^M P_{\ell,m}(\gamma_{SR}) P_{m,\ell}(\gamma_{RD}). \quad (15)$$

On taking expectation of the conditional probability in (15), the end-to-end average probability of correct detection is expressed as

$$P_c = \frac{1}{M} \sum_{\ell=1}^M \sum_{m=1}^M P_{\ell,m}^{SR} P_{m,\ell}^{RD}, \quad (16)$$

where $P_{\ell,m}^{SR} = E[P_{\ell,m}(\gamma_{SR})]$ and $P_{m,\ell}^{RD} = E[P_{m,\ell}(\gamma_{RD})]$ are the average error probabilities. Let Ψ be a set of SR and RD links, that is, $\Psi = \{SR, RD\}$. In signaling schemes with symmetric constellations, the paired error probability is dependent on the distance between the two constellation points, therefore $P_{\ell,m}^\varphi = P_{m,\ell}^\varphi$, $\varphi \in \Psi$. We represent $P_{\ell,m}^\varphi = P_{m,\ell}^\varphi = P_{|\ell-m|}^\varphi$, for $\ell, m \in \{1, \dots, M\}$. Thus, (16) can be re-written as

$$\begin{aligned} P_c &= \frac{1}{M} \sum_{\ell=1}^M \sum_{m=1}^M P_{|\ell-m|}^{SR} P_{|\ell-m|}^{RD} \\ &= \sum_{\nu=0}^{M-1} P_\nu^{SR} P_\nu^{RD}. \end{aligned} \quad (17)$$

The average probabilities of correct decision corresponding to $\nu = 0$ in (17) are represented as $P_0^{SR} = (1 - P_{e,SR})$ and $P_0^{RD} = (1 - P_{e,RD})$. $P_{e,SR}$ and $P_{e,RD}$ are the average error probabilities for SR and RD links, respectively. Thus, end-to-end probability of correct decision can be represented in terms of the average link error probabilities and the pairwise error probabilities as

$$P_c = (1 - P_{e,SR})(1 - P_{e,RD}) + \sum_{\nu=1}^{M-1} P_\nu^{SR} P_\nu^{RD}. \quad (18)$$

Using (18), the end-to-end average SER is given by

$$P_e = P_{e,SR} + P_{e,RD} - P_{e,SR} P_{e,RD} - \sum_{\nu=1}^{M-1} P_\nu^{SR} P_\nu^{RD}. \quad (19)$$

A. Analytical Average SER

1) M -PSK: The average error terms in (19) are represented in integral form using [18, eqs. (7) and (8)] as

$$P_{e,\varphi} = \frac{1}{\pi} \int_0^{\phi_0} M_{\gamma_\varphi} \left(\frac{g_0}{\sin^2 \theta} \right) d\theta \quad \text{and} \quad (20)$$

$$\begin{aligned} P_\ell^\varphi &= \frac{1}{2\pi} \int_0^{\phi_1} M_{\gamma_\varphi} \left(\frac{g_1}{\sin^2 \theta} \right) d\theta \\ &\quad - \frac{1}{2\pi} \int_0^{\phi_2} M_{\gamma_\varphi} \left(\frac{g_2}{\sin^2 \theta} \right) d\theta, \quad \varphi \in \Psi, \end{aligned} \quad (21)$$

where $\phi_0 = (M-1)\pi/M$, $g_0 = \sin^2(\pi - \phi_0)$, $\phi_1 = (\pi - 2\pi\ell/M + \pi/M)$, $g_1 = \sin^2(\pi - \phi_1)$, $\phi_2 = (\pi - 2\pi\ell/M - \pi/M)$, $g_2 = \sin^2(\pi - \phi_2)$, and $M_{\gamma_\varphi}(s)$ is the moment generating function (MGF) of γ_φ . The MGF of γ_φ is defined as

$$M_{\gamma_\varphi}(s) = \int_0^\infty \exp(-s\gamma_\varphi) f_{\gamma_\varphi}(\gamma_\varphi) d\gamma_\varphi, \quad \varphi \in \Psi, \quad (22)$$

where $f_{\gamma_\varphi}(\gamma_\varphi)$ is PDF of γ_φ . The PDF of γ_{SR} can be obtained using [17, eq. (A.5)] as

$$f_{\gamma_{SR}}(\gamma_{SR}) = \frac{2}{\gamma_{SR}} K_0 \left(2 \sqrt{\frac{\gamma_{SR}}{\bar{\gamma}_{SR}}} \right), \quad (23)$$

where $K_0(\cdot)$ is the 0th order modified Bessel's function of the second kind. Substituting (23) in (22) and using [20, eq. (2.16.8.4)], the MGF of γ_{SR} is given by

$$M_{\gamma_{SR}}(s) = \left(\frac{1}{s\bar{\gamma}_{SR}} \right)^{1/2} \exp \left(\frac{1}{2s\bar{\gamma}_{SR}} \right) W_{-\frac{1}{2},0} \left(\frac{1}{s\bar{\gamma}_{SR}} \right), \quad (24)$$

where $W_{s,t}(\cdot)$ is Whittaker's function. The PDF of γ_{RD} can be obtained using (12b) and (13), which when substituted in (22) gives the corresponding MGF as

$$M_{\gamma_{RD}}(s) = (1 + s\bar{\gamma}_{RD})^{-1}. \quad (25)$$

The expression for MGF in (25) is a well known result.

On substituting (20) and (21) in (19) and using (24) and (25), the end-to-end average SER for M -PSK modulation scheme can be obtained.

2) Orthogonal M -FSK: In case of orthogonal M -FSK, all constellation points are at equal distance, therefore the paired error probabilities are related as $P_1^\varphi = \dots = P_{M-1}^\varphi$, $\varphi \in \Psi$ [19]. Therefore, the end-to-end average probability of error in (19) can also be written as

$$P_e = P_{e,SR} + P_{e,RD} - P_{e,SR}P_{e,RD} - (M-1)P_1^{SR}P_1^{RD}, \quad (26)$$

where $P_1^{SR} = P_{e,SR}/(M-1)$ and $P_1^{RD} = P_{e,RD}/(M-1)$. On further simplifying (26), we get

$$P_e = P_{e,SR} + P_{e,RD} - \frac{M}{M-1}P_{e,SR}P_{e,RD}. \quad (27)$$

Next, using the well known expression for the conditional error probability of orthogonal M -FSK with non-coherent detection, the conditional error probability for link φ is given by [21, eq. (13.59c)]

$$P_{e,\varphi}(\gamma_\varphi) = \sum_{l=1}^{M-1} \frac{(-1)^{l+1}}{(l+1)} \binom{M-1}{l} \exp \left(-\frac{l\gamma_\varphi}{(l+1)} \right). \quad (28)$$

The average of error probability $P_{e,SR}$ can be found by taking expectation of (28) using the PDF in (23), which on simplification using [20, eq. (2.16.8.4)] results in

$$\begin{aligned} P_{e,SR} &= \sum_{l=1}^{M-1} \frac{(-1)^{l+1}}{(l+1)} \binom{M-1}{l} \left(\frac{(l+1)}{l\bar{\gamma}_{SR}} \right)^{1/2} \\ &\quad \times \exp \left(\frac{(l+1)}{2l\bar{\gamma}_{SR}} \right) W_{-\frac{1}{2},0} \left(\frac{(l+1)}{l\bar{\gamma}_{SR}} \right). \end{aligned} \quad (29)$$

Next, $P_{e,RD}$ in (27) can be obtained using (28), the PDF of γ_{RD} defined in terms of (12b) and (13), and the relation $\int_0^\infty z^{s-1} \exp(-tz) dz = t^{-s} \Gamma(s)$ as

$$P_{e,RD} = \sum_{l=1}^{M-1} \frac{(-1)^{l+1}}{(l+1)} \binom{M-1}{l} \left(1 + \frac{l\bar{\gamma}_{RD}}{(l+1)} \right)^{-1}. \quad (30)$$

On substituting (29) and (30) in (27), the end-to-end average SER for orthogonal M -FSK can be obtained.

B. Asymptotic Average SER

At high SNRs, the end-to-end average SER expression in (19) can be approximated as

$$P_e^\infty \approx P_{e,SR}^\infty + P_{e,RD}^\infty, \quad (31)$$

where P_e^∞ , $P_{e,SR}^\infty$, and $P_{e,RD}^\infty$ are asymptotic approximation of P_e , $P_{e,SR}$, and $P_{e,RD}$, respectively.

1) M -PSK: The high SNR approximation of P_e in (31) can be analyzed by determining the asymptotic approximation of $P_{e,SR}$ and $P_{e,RD}$. $P_{e,SR}^\infty$ can be obtained by approximating $f_{\gamma_{SR}}(\gamma_{SR})$ using [22, eq. (9.6.8)]. The PDF expression thus obtained is then used to analyze the corresponding MGF $M_{\gamma_{SR}}^\infty(s)$ and the error term $P_{e,SR}^\infty$ as follows.

The PDF in (23) is approximated as

$$f_{\gamma_{SR}}^\infty(\gamma_{SR}) \approx -\frac{2}{\bar{\gamma}_{SR}} \ln \left(\sqrt{\frac{4\gamma_{SR}}{\bar{\gamma}_{SR}}} \right). \quad (32)$$

The corresponding MGF can be obtained by substituting (32) in (22) and using [23, eq. (4.352.1)]. That is

$$M_{\gamma_{SR}}^\infty(s) \approx \left(\ln \left(\frac{s\bar{\gamma}_{SR}}{4} \right) - \psi(1) \right) \frac{1}{s\bar{\gamma}_{SR}}, \quad (33)$$

where $\psi(\cdot)$ is digamma function [22, eq. (6.3.1)]. On replacing (33) in (20), the approximation of $P_{e,SR}^\infty$ is given by

$$P_{e,SR}^\infty \approx \frac{1}{\pi g_0 \bar{\gamma}_{SR}} \left(\left(\ln \left(\frac{g_0 \bar{\gamma}_{SR}}{4} \right) - \psi(1) \right) \mathcal{I} + \mathcal{J} \right), \quad (34)$$

where

$$\mathcal{I} = \int_0^{\phi_0} \sin^2(\theta) d\theta, \quad \text{and} \quad (35)$$

$$\mathcal{J} = \int_0^{\phi_0} \sin^2(\theta) \ln(\sin^2(\theta)) d\theta. \quad (36)$$

Now, in order to obtain $P_{e,RD}^\infty$ in (31), the asymptotic approximation of MGF $M_{\gamma_{RD}}(s)$ is analyzed. The MGF in (25) for $(s\bar{\gamma}_{RD}) \gg 1$ is

$$M_{\gamma_{RD}}^\infty(s) \approx (s\bar{\gamma}_{RD})^{-1}. \quad (37)$$

Substituting (37) in (20), $P_{e,RD}$ is rewritten as

$$P_{e,RD}^\infty \approx (\pi g_0 \bar{\gamma}_{RD})^{-1} \mathcal{I}. \quad (38)$$

On substituting (34) and (38), the asymptotic end-to-end average SER for M -PSK in (31) is expressed as

$$P_e^\infty \approx \frac{\mathcal{A}((\mathcal{B} + \ln(\bar{\gamma}_{SR})) \mathcal{I} + \mathcal{J})}{\bar{\gamma}_{SR}} + \frac{\mathcal{C}\mathcal{I}}{\bar{\gamma}_{RD}}, \quad (39)$$

where $\mathcal{A} = 1/(g_0\pi)$, $\mathcal{B} = \ln(g_0/4) - \psi(1)$, and $\mathcal{C} = 1/(g_0\pi)$.

Following the relation in [4, eq. (22)], integrals (35) and (36) can be approximated as $\mathcal{I} \approx a_M \int_0^{\pi/2} \sin^2(\theta) d\theta$ and $\mathcal{J} \approx b_M \int_0^{\pi/2} \sin^2(\theta) \ln(\sin^2(\theta)) d\theta$, respectively, where $a_M, b_M = 1$ for $M = 2$ (binary phase-shift keying (BPSK)) and $a_M, b_M = 2$ for $M \geq 4$. The approximation is exact for $M = 2$ and close for $M \geq 4$. Using [23, eqs. (3.621.1) and (4.387.2)], the closed-form representation of the approximations for \mathcal{I} and \mathcal{J} are $\mathcal{I} \approx a_M \pi/4$ and $\mathcal{J} \approx b_M \pi(\psi(3/2) - \psi(2))/4$, respectively.

2) Orthogonal M -FSK: The high SNR approximation of $P_{e,SR}^\infty$ for orthogonal M -FSK can be obtained by averaging the conditional SER in (28) using the approximate expression of $f_{\gamma_{SR}}(\gamma_{SR})$. The resulting $P_{e,SR}^\infty$ is determined by taking expectation of (28) using (32) and [23, eq. (4.253.1)] as

$$P_{e,SR}^\infty \approx \frac{1}{\bar{\gamma}_{SR}} \sum_{l=1}^{M-1} \mathcal{D}(l) \left(\ln \left(\frac{l\bar{\gamma}_{SR}}{4(l+1)} \right) - \psi(1) \right), \quad (40)$$

where

$$\mathcal{D}(l) = \frac{(-1)^{l+1}}{(l+1)} \binom{M-1}{l} \left(\frac{l+1}{l} \right). \quad (41)$$

At high SNRs, $l\bar{\gamma}_{RD}/((l+1)) \gg 1$ in (30), thus the asymptotic approximation of $P_{e,RD}$ is

$$P_{e,RD}^{\infty} \approx \frac{1}{\bar{\gamma}_{RD}} \sum_{l=1}^{M-1} \mathcal{D}(l). \quad (42)$$

The asymptotic end-to-end average SER for orthogonal M -FSK in (31) can be expressed using (40) and (42) as

$$P_e^{\infty} \approx \frac{\mathcal{A}(l; M) + \mathcal{B}(l; M) \ln(\bar{\gamma}_{SR})}{\bar{\gamma}_{SR}} + \frac{\mathcal{C}(l; M)}{\bar{\gamma}_{RD}}, \quad (43)$$

where

$$\mathcal{A}(l; M) = \sum_{l=1}^{M-1} \mathcal{D}(l) \left(\ln \left(\frac{l}{4(l+1)} \right) - \psi(1) \right)$$

$$\mathcal{B}(l; M) = \sum_{l=1}^{M-1} \mathcal{D}(l), \quad \text{and} \quad \mathcal{C}(l; M) = \sum_{l=1}^{M-1} \mathcal{D}(l).$$

C. Optimal Relay Location

The optimal relay location for M -PSK and orthogonal M -FSK modulation schemes can be obtained using (39) and (43), respectively. Using (14a) and (14b), (39) can be rewritten as

$$P_e^{\infty} \approx \mathcal{A}'((\mathcal{B}' - (\alpha_{RS} + \alpha_{SR}) \ln(d_{SR})) \mathcal{I} + \mathcal{J}) \\ \times (d_{SR})^{(\alpha_{RS} + \alpha_{SR})} + \mathcal{C}'(\beta d_{SD} - d_{SR})^{\alpha_{RD}} \mathcal{I}, \quad (44)$$

where $\mathcal{A}' = \mathcal{A}\mathcal{Z}_1$, $\mathcal{B}' = \mathcal{B} - \rho_1 \ln(\mathcal{Z}_1)$, and $\mathcal{C}' = \mathcal{C}\mathcal{Z}_2$, with $\mathcal{Z}_1 = 1/(\eta\lambda_{RS}\lambda_{SR}\bar{\gamma})$, $\mathcal{Z}_2 = 1/(\lambda_{RD}\bar{\gamma})$, and $\bar{\gamma} = P_R T_s / N_0$. We consider that the nodes are planner and their inter link distances follow the relation $d_{SR} + d_{RD} = \beta d_{SD}$, $\beta \geq 1$. For $\beta = 1$, the nodes are collinear. In case of orthogonal M -FSK, (43) is rewritten as

$$P_e^{\infty} \approx (\mathcal{A}(l; M) - \mathcal{B}(l; M)(\ln(\mathcal{Z}_1) + (\alpha_{RS} + \alpha_{SR}) \ln(d_{SR}))) \\ \times \mathcal{Z}_1 (d_{SR})^{(\alpha_{RS} + \alpha_{SR})} + \mathcal{C}(l; M) \mathcal{Z}_2 (\beta d_{SD} - d_{SR})^{\alpha_{RD}}. \quad (45)$$

On applying the second-order conditions [24], it is identified that (44) and (45) are convex functions of d_{SR} . Therefore, at high SNRs an optimal relay location achieving minimum average SER for M -PSK and orthogonal M -FSK can be obtained using (44) and (45), respectively. Golden-section search method [4] is used to obtain the optimal values.

D. Diversity Order

The asymptotic average SER in (39) and (43) can be rewritten in the form

$$P_e^{\infty} = \mathcal{Y}_1(\mathcal{Y}_2 + \ln(\bar{\gamma}))/\bar{\gamma} + \mathcal{Y}_3/\bar{\gamma}, \quad (46)$$

where \mathcal{Y}_1 , \mathcal{Y}_2 , and \mathcal{Y}_3 represent the terms independent of $\bar{\gamma}$ ($= P_R T_s / N_0$). \mathcal{Y}_1 , \mathcal{Y}_2 , and \mathcal{Y}_3 are different for the two-modulation schemes considered in this paper. Now, using (46) and the relation to obtain the DO of a system $\text{DO} = -\lim_{\bar{\gamma} \rightarrow \infty} \ln(P_e^{\infty})/\ln(\bar{\gamma})$, we get

$$\text{DO} = \min \{(1 - \ln \ln(\bar{\gamma})/\ln(\bar{\gamma})), 1\}. \quad (47)$$

At significantly high SNR DO tends to 1.

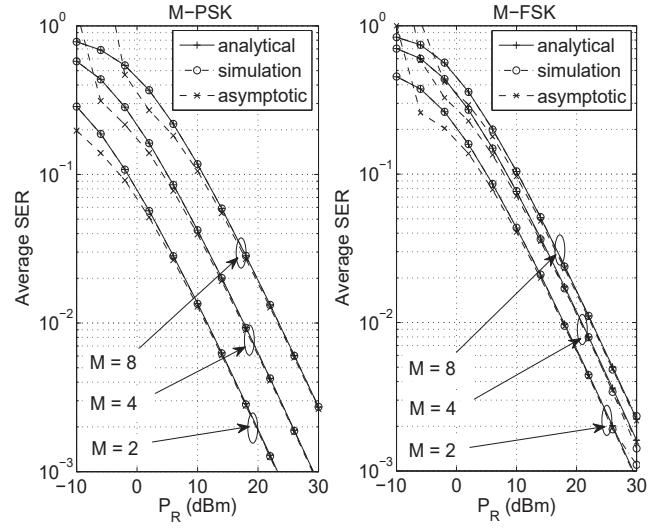


Fig. 2. Average SER versus transmission power P_R for M -PSK and orthogonal M -FSK when $m_{ij} = 1$.

IV. NUMERICAL RESULTS AND DISCUSSION

In this section, the end-to-end average SER derived in Section III-A and the corresponding asymptotic average SER in Section III-B are presented. We consider energy conversion unit at the source node operates optimally with maximum efficiency, thus $\eta = 1$. Without loss of generality, we also assume that symbol duration $T_s = 1$ second. The numerical results are plotted on considering the noise PSD $N_0 = 10^{-4}$ and the transmit power is measured in dBm.

The average SER for M -PSK and orthogonal M -FSK are plotted with power transmitted by the relay node for different modulation index M in Fig. 2. For M -PSK, the average SER (P_e) is plotted using (19) on substituting (20), (21), (24), and (25), whereas for orthogonal M -FSK the terms in (27) are evaluated using (29) and (30). Channel gains are assumed to be unity, that is, $\lambda_{ij}(d_{ij})^{-\alpha_{ij}} = 1$. Simulation results are also plotted in Fig. 2 to validate the analysis presented in Section III-A. We observe from Fig. 2 that the average SER raises with increase in the modulation order. Further, plots for the asymptotic average SER are also shown in Fig. 2. The expressions for the asymptotic average SER are found in Section III-B for M -PSK and orthogonal M -FSK schemes as (39) and (43), respectively. At high SNRs, the asymptotic results are found to be a close approximation of the average SER. The numerical results can be useful for selecting the modulation scheme and the modulation order while ensuring reliability, feasible system complexity, and resources available such as power and bandwidth, etc.

In Fig. 3, plots showing variation in the average SER with source-to-relay distance d_{SR} are presented for 4-PSK and orthogonal 4-FSK schemes, respectively. Nodes are considered to be coplanar and inter-link distances follow the relation $d_{SR} + d_{RD} = \beta d_{SD}$. We assume $\beta = 1.1$, $d_{SD} = 2$ unit, $\lambda_{ij} = 1$, and $\alpha_{ij} = 3$ for the analysis. We observe that average SER for both modulation schemes is a convex

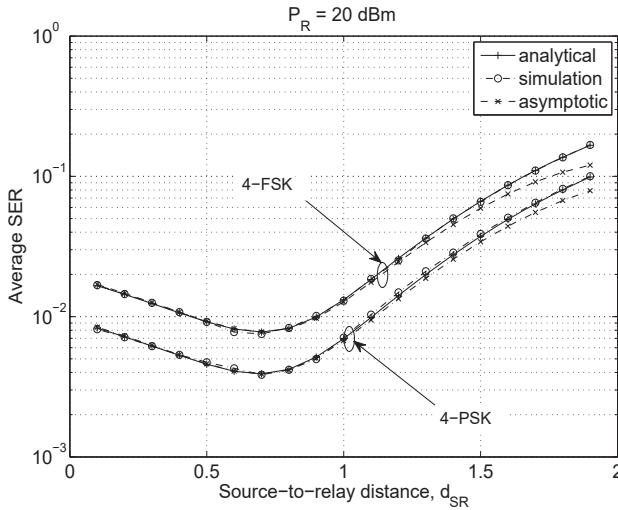


Fig. 3. Variation in average SER with source-to-relay distance d_{SR} for $d_{SR} + d_{RD} = \beta d_{SD}$, $\beta = 1.1$, $d_{SD} = 2$ unit, $\lambda_{ij} = 1$, $\alpha_{ij} = 3$.

function of the distance d_{SR} . Thus, Golden-section search method can be applied to minimize asymptotic average SER in (44) for M -PSK and (45) for orthogonal M -FSK; the optimal relay locations for the two modulation schemes are given by 0.6970 and 0.7067, respectively. The results are also found for different modulation orders and power at the relay node. Observation is made that for the two modulation schemes optimal relay location is always close to the source ($d_{SR} \approx 0.7$).

V. CONCLUSION

We analyze the end-to-end average SER for a WP two-hop DF relay system where the source node is powered by the relay node. The direct link between source and destination is considered to be deeply faded and therefore ignored for the analysis. Other links are modeled using Rayleigh distribution. The analysis is presented when transmitted data is i) M -PSK modulated with coherent detection and ii) orthogonal M -FSK modulated with noncoherent detection. Simulation results for the average SER are plotted along with the numerical results and they are in perfect agreement. The high SNR approximations of the average SER are also obtained for the two modulation schemes. Further, using the asymptotic approximation of the average SER, optimal relay location is identified. In this paper, the observations made are summarized as follows: i) average SER for M -PSK with coherent detection is less than that of the orthogonal M -FSK with non-coherent detection for same modulation order ii) average SER increases with increment in the modulation order for the two schemes iii) variation in average SER with the modulation order is more for M -PSK than M -FSK, and iv) the optimal relay location is close to the source node irrespective of the transmission power and modulation scheme/order.

REFERENCES

- [1] S. Y. Lee, T. Y. Chen, C. Tsou, and Y. S. Chu, "Wireless energy harvesting circuit and system with error-correction ASK demodulator for body sensor network with ultra-high-frequency RFID healthcare system," *IET Wireless Sensor Syst.*, vol. 8, no. 1, pp. 36–44, 2018.
- [2] Y. Alsaba, S. K. A. Rahim, and C. Y. Leow, "Beamforming in wireless energy harvesting communications systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 1329–1360, 2nd-quarter 2018.
- [3] H. Zhang and W. X. Zheng, "Robust transmission power management for remote state estimation with wireless energy harvesting," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2682–2690, Aug. 2018.
- [4] P. Kumar and K. Dhaka, "Performance analysis of wireless powered DF relay system under Nakagami- m fading," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7073–7085, Aug. 2018.
- [5] S. Sudevalayam and P. Kulkarni, "Energy harvesting sensor nodes: Survey and implications," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 443–461, 3rd-quarter 2011.
- [6] A. A. Nasir, X. Zhou, S. Durrani, and R. A. Kennedy, "Relaying protocols for wireless energy harvesting and information processing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3622–3636, Jul. 2013.
- [7] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, May 2013.
- [8] P. Liu, S. Gazor, I. Kim, and D. Kim, "Energy harvesting noncoherent cooperative communications," *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6722–6737, Dec. 2015.
- [9] F. Zhao, L. Wei, and H. Chen, "Optimal time allocation for wireless information and power transfer in wireless powered communication systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1830–1835, Mar. 2016.
- [10] K. J. R. Liu, A. K. Sadek, W. Su, and A. Kwasinski, *Cooperative Communications and Networking*. Cambridge Univ. Press, 2009.
- [11] D. Mishra, S. De, and C. F. Chiasserini, "Joint optimization schemes for cooperative wireless information and power transfer over Rician channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 554–571, Feb. 2016.
- [12] E. Chen, M. Xia, D. B. da Costa, and S. Aïssa, "Multi-hop cooperative relaying with energy harvesting from cochannel interferences," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1199–1202, May 2017.
- [13] H. Chen, X. Zhou, Y. Li, P. Wang, and B. Vucetic, "Wireless-powered cooperative communications via a hybrid relay," in *Proc. of the IEEE Inform. Theory Workshop*, Nov. 2014, pp. 666–670.
- [14] N. Zlatanov, D. W. K. Ng, and R. Schober, "Capacity of the two-hop relay channel with wireless energy transfer from relay to source and energy transmission cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 647–662, Jan. 2017.
- [15] Z. Yang, W. Xu, Y. Pan, C. Pan, and M. Chen, "Energy efficient resource allocation in machine-to-machine communications with multiple access and energy harvesting for IoT," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 229–245, Feb. 2018.
- [16] S. Luo, G. Yang, and K. C. Teh, "Throughput of wireless-powered relaying systems with buffer-aided hybrid relay," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4790–4801, Jul. 2016.
- [17] A. A. Nasir, X. Zhou, S. Durrani, and R. A. Kennedy, "Throughput and ergodic capacity of wireless energy harvesting based DF relaying network," in *Proc. of the IEEE Int. Conf. on Commun.*, Jun. 10–14, 2014, pp. 4066–4071.
- [18] M. D. Selvaraj and R. K. Mallik, "Error analysis of the decode and forward protocol with selection combining," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3086–3094, Jun. 2009.
- [19] K. Dhaka, R. K. Mallik, and R. Schober, "Performance analysis of decode-and-forward multi-hop communication: A difference equation approach," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 339–345, Feb. 2012.
- [20] A. P. Prudnikov, Y. A. Brychkov, O. I. Marichev, and G. G. Gould, *Integrals and Series: Special Functions*. New York, NY, USA:Gordon and Breach Science, 1992, vol. 2.
- [21] B. P. Lathi, *Modern Digital and Analog Communication Systems*. 3rd ed. New York, NY, USA: Oxford Uni. Press, 1998.
- [22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington: U.S. Govt. Print. Off., 1970.
- [23] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. 6th ed. New York, NY, USA: Academic, 2000.
- [24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge Univ. Press, 2004.

Performance Evaluation of Visible Light Communication for DCO and ACO Optical OFDM Techniques

Mahendra P. S. Bhadoria^{a*}, Gaurav Pandey^{b†}, Abhishek Dixit^{a,b‡}

^aBharti School of Telecommunication Technology and Management, Indian Institute of Technology (IIT), Delhi, New Delhi, India

^bDepartment of Electrical Engineering, Indian Institute of Technology (IIT), Delhi, New Delhi, India

Email: *mahendrabhadoria892@gmail.com, †gauravpandey@iitd.ac.in, ‡abhishek.dixit@iitd.ac.in

Abstract—Visible light communication (VLC) has evolved as a relatively new research topic that employs white light-emitting diodes for data transmission. The basic requirement for optical transmission is that signal should be real and positive. Among the various developed modulation techniques for VLC, orthogonal frequency division multiplexing (OFDM) has drawn major consideration because of its high data rate and heftiness to inter-symbol interference (ISI), but it suffers from the problem of high peak to average power ratio (PAPR), causing signal distortion thereby effecting the system efficiency. In this paper, asymmetrically clipped optical OFDM (ACO-OFDM) and direct current biased optical OFDM (DCO-OFDM) for VLC have been investigated by evaluating its bit error rate (BER) and PAPR performance. The effect of the uncorrelated noise due to dual-sided clipping of the signal, modulation order and number of subcarriers on the symbol distortion is also studied. Study on selection of optimum clipping levels is done to reduce the PAPR for both the schemes. Simulation analysis implies that ACO-OFDM is better than DCO-OFDM by around 4.5 dB for a BER of 1×10^{-3} and by about 2 dB for a PAPR complementary cumulative distribution function (CCDF) of 1×10^{-1} .

Index Terms—Visible Light Communication (VLC), orthogonal frequency division multiplexing (OFDM), peak to average power ratio (PAPR), ACO-OFDM, DCO-OFDM.

I. INTRODUCTION

The current scenario in the field of wireless communication in terms of the increasing need for greater throughput and worldwide coverage area have a great requirement for efficient use of the spectrum of radio frequency (RF) communication. The number of devices accessing wireless networks has increased drastically over the past few years thereby increasing mobile data traffic. Apart from this, wireless data traffic is further increased by the growing trend of online social services. Other factors that cause the problem in RF communication are security, interference, health safety, and power inefficiency. All these factors call for some alternative means for accommodating extra capacity in the future. One of the most effective alternative in the present day is the use of light-emitting diodes (LED) technology in wireless visible light communication (VLC). LEDs have an advantage of providing large modulation bandwidth along with energy efficient lighting.

Visible light provides around 400 THz bandwidth for secure and unlicensed wireless communication. Its bandwidth

is more than thousand times broader than presently used RF that enables the great potential for data and communication purposes [1].

Proper standardization for VLC was established in 2011 when the Institute of Electrical and Electronics Engineers (IEEE) proposed IEEE 802.11.7 standards [2]. The main idea to use VLC is to combine informative data with the LED light changes and then detecting at the receiving end known as intensity-modulation and direct-detection (IM/DD). IM/DD technique is used for data transmission on optical wireless channels. In the case of the various single-carrier pulse modulation schemes, the major limiting factor for data rate is the dispersion induced by optical wireless channel due to inter-symbol interference (ISI) [3]. Therefore, multi-carrier modulation evolved which have robustness to ISI. Multilevel quadrature amplitude modulation (M-QAM) enabled optical orthogonal frequency division multiplexing (O-OFDM) have the capability to provide higher data rates [4]. In the case of O-OFDM, the LED intensity is modulated by the time-domain envelope. The frequency which is used for modulation is higher than the frequency that is detected by the human eyes and thus it cannot cause noticeable flickering. Apart from this, vast research is done on modulation and demodulation techniques of visible light to improve the data rates [5, 6].

VLC provides various advantages over RF communications [7] including interference protection with RF signals, secure communication, no health hazards, energy efficient data transmission, easy implementation using existing infrastructure and low cost. As compared to the RF communication, transmission using visible light is cheaper. Visible light communication has a large number of applications including underwater communication, vehicular communication, indoor positioning systems and light-fidelity (Li-Fi).

In this paper, direct current biased optical OFDM (DCO-OFDM) and asymmetrically clipped optical OFDM (ACO-OFDM) techniques for VLC are analyzed by evaluating its peak to average power ratio (PAPR) and bit error rate (BER) performance. The basic requirement for optical transmission is that the signal should be real and positive. Apart from this, OFDM is responsible for high data rate and reduced ISI but it suffers from the problem of high PAPR which causes signal

distortion thereby effecting the system efficiency. Furthermore, for reducing the problem of large PAPR, a detailed analysis on wise selection of optimum clipping levels for both the schemes has been done.

II. VLC SYSTEM MODEL BASED ON OPTICAL OFDM

Figure 1 shows the system model for VLC based on optical OFDM. OFDM is highly used in the wireless and wired communication due to its ability to oppose ISI. It brings with it additional merits too other than ISI resistance such as better optical power and spectral efficiency, higher data rate, etc.

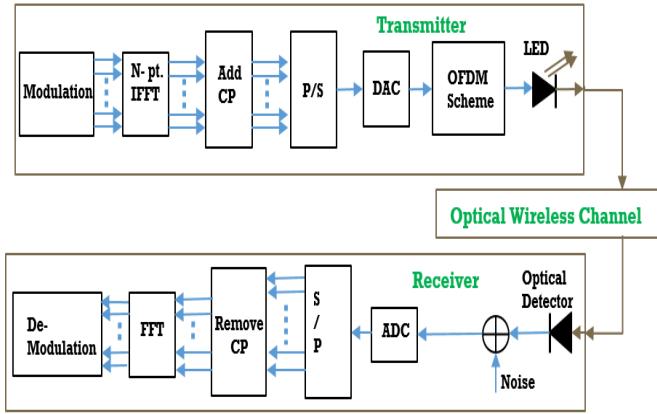


Fig. 1: System model for VLC based on optical OFDM.

In general OFDM, the transmitted signal is complex and bipolar in nature but directly using it as such in VLC systems using IM/DD is not possible as the signals are transmitted in form of optical power, necessitating it to be non-negative. To get a real valued signal, OFDM subcarriers are imposed with Hermitian symmetry and to make the signal non-negative, addition of direct current (DC) bias in DCO-OFDM and signal clipping at zero is done in ACO-OFDM. In ACO-OFDM, data transmission occurs on only odd subcarriers. Since, only even subcarriers (having no information) are affected by the clipping, thus data transmission occurs without any information loss.

Initially, the information bits are mapped to frequency domain symbols depending on the employed modulation technique e.g. BPSK, M-QAM, etc. The transformation of the frequency domain signal to the time domain is done by employing inverse fast Fourier transform (IFFT). For a F size IFFT, N subcarriers are used for implementing OFDM scheme. Here, $N = F$ is used for simulations. In ACO-OFDM, the first half odd subcarriers are modulated by the first $N/4$ symbols and in DCO-OFDM, the first $N/2 - 1$ subcarriers by the first $N/2 - 1$ symbols. The other left half subcarriers are applied with Hermitian symmetry in both systems to make sure that the signal obtained after IFFT is real-valued. Thus, the frequency domain modulated signal, $Y = [Y_0, Y_1, Y_2, \dots, Y_{N-1}]$ satisfy the following conditions,

$$Y_0 = Y_{N/2} = 0 \quad (1)$$

$$Y_J = Y_{N-J}^*, J = 1, 2, 3, \dots, N-1 \quad (2)$$

where even N are not modulated in case of ACO-OFDM. After IFFT, signal obtained can be expressed as,

$$y_m = \frac{1}{N} \sum_{J=0}^{N-1} Y_J \exp\left(j \frac{2\pi Y m}{N}\right) \quad (3)$$

The above equation can be expressed as,

$$y_m = \frac{2}{N} \sum_{J=0}^{\frac{N}{2}-1} \Re\left(Y_J \exp\left(j \frac{2\pi Y m}{N}\right)\right) \quad (4)$$

where J^{th} subcarrier of Y is Y_J .

Due to $Y_0 = Y_{N/2} = 0$ and Hermitian symmetry, the subcarriers with relevant information are only $(N/2 - 1)$.

In case of ACO-OFDM, this can be further simplified as,

$$y_m = \frac{2}{N} \sum_{J=1}^{\frac{N}{4}} \Re\left(Y_{2J-1} \exp\left(j \frac{2\pi(2J-1)m}{N}\right)\right) \quad (5)$$

$$y_{m+\frac{N}{2}} = \frac{2}{N} \sum_{J=1}^{\frac{N}{4}} \Re\left(Y_{2J-1}^* \exp\left(j \frac{2\pi(2J-1)(m+\frac{N}{2})}{N}\right)\right) = -y_m \quad (6)$$

So, the clipping can be done without any loss in the transmitted information. After this, the real signal is transformed from parallel to serial, then cyclic-prefix is combined with the signal to decrease ISI. The resulting signal is then passed through a digital to analog converter (DAC) and low pass filter. The signal obtained after this is real in nature, but to make it positive, addition of DC bias is done and the residual negative peaks are removed at 0 level.

As can be seen, ACO-OFDM uses only half sub-carriers, it shows a decrement in the spectral efficiency by 50% as compared to DCO-OFDM but has better optical power efficiency. Since in DCO-OFDM, DC bias is carrying no information and constitutes as the major component in the optical power transmitted.

Spectral efficiency (η) of the DCO-OFDM and ACO-OFDM [8] is written as,

$$\eta_{DCO} (\text{ bits/sec/Hz}) = \frac{N-2}{2(N+N_g)} \log_2(M) \quad (7)$$

$$\eta_{ACO} (\text{ bits/sec/Hz}) = \frac{N}{4(N+N_g)} \log_2(M) \quad (8)$$

where, N_g is the number of the symbols added as cyclic prefix.

OFDM signals have an issue of high PAPR since high bias is needed to make the signal positive. Thus, instead of adding a very high bias, a medium level bias is given and the residual negative values are clipped at 0 causing clipping noise. The data rate R of the ACO-OFDM and DCO-OFDM can be written as,

$$R_{ACO} = \frac{\frac{N}{4} - 1}{N + N_g} B \log_2(M) \text{ bits/s} \quad (9)$$

$$R_{DCO} = \frac{\frac{N}{2} - 1}{N + N_g} B \log_2(M) \text{ bits/s} \quad (10)$$

For large N , $y(t)$ has Gaussian distribution with 0 mean i.e. $E[y(m)] = 0$ and σ_y^2 variance i.e. $E[y(m)^2]$ (Both the systems will be having different distributions as both have different frame structures).

Signal can also be clipped from both ends to keep the signal in the transmitter dynamic range and to reduce the PAPR. This clipping from both ends (i.e. λ_{bottom} (normalized bottom clipping level) and λ_{top} (normalized top clipping level) causes clipping noise which affects both the BER and PAPR performance. It is done as,

$$y_{clip} = \begin{cases} H_u & ; y_m \geq H_u \\ y_m & ; -H_l < y_m \leq H_u \\ -H_l & ; y_m \leq -H_l \end{cases} \quad (11)$$

For ACO-OFDM, $H_l = 0$

$$\lambda_{bottom} = \frac{-H_l}{\sigma_y} \quad \text{and} \quad \lambda_{top} = \frac{H_u}{\sigma_y} \quad (12)$$

The probability distribution function (PDF) of $y_{DCO}(t)$ is clipped Gaussian distribution given by,

$$f_{DCO}(\omega) = \phi(\lambda_{bottom})\delta(\omega + H_l) + \frac{1}{\sqrt{2\pi}\sigma_y^2} \exp\left(-\frac{\omega^2}{2\sigma_y^2}\right) [u(\omega + H_l) - u(\omega - H_l)] + (1 - \phi(\lambda_{bottom}))\delta(\omega - H_u) \quad (13)$$

where,

$$\phi(p) = \int_{-\infty}^p \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt \quad (14)$$

The PDF of ACO-OFDM can be computed just by keeping $H_l = 0$. After clipping distortion, signal attenuation (K) occurs along with addition of non-Gaussian uncorrelated clipping noise which is bipolar in case of DCO-OFDM and unipolar in case of ACO-OFDM. Following the procedures elaborated in [9], these can be written as,

$$K = Q(\lambda_{bottom}) - Q(\lambda_{top}) \quad (15)$$

where, $Q(\cdot)$ is the Q-function of normal distribution.

In ACO-OFDM, clipping noise variance is given by,

$$\sigma_{clip}^2 = \frac{\sigma_y^2}{2}(1 - 4K^2) - \Delta\sigma^2 \quad (16)$$

where,

$$\begin{aligned} \Delta\sigma^2 = & \frac{\sigma_y^2}{2}[1 - (4\phi(\lambda_{top})\lambda_{bottom} - 2\phi(\lambda_{bottom}) \\ & \lambda_{bottom} - 2\phi(\lambda_{top})\lambda_{top} + (\lambda_{bottom}^2 + 1) \\ & (1 - 2Q(\lambda_{top})) - (\lambda_{bottom}^2 + 1)(1 - 2Q(\lambda_{bottom})) \\ & + 2Q(\lambda_{top})\lambda_{top}(\lambda_{top} - \lambda_{bottom})^2)] \end{aligned} \quad (17)$$

In DCO-OFDM, variance is given by,

$$\sigma_{clip}^2 = \frac{\sigma_y^2}{2}(1 - K^2) - \Delta\sigma^2 \quad (18)$$

where,

$$\begin{aligned} \Delta\sigma^2 = & \sigma_y^2 - \sqrt{\sigma_y^2}(\phi(\lambda_{bottom}) - \phi(\lambda_{top}) + (1 - Q(\lambda_{bottom})) \\ & \lambda_{bottom} + Q(\lambda_{top})\lambda_{top}) + \sigma_y^2(Q(\lambda_{bottom}) - Q(\lambda_{top}) \\ & + \phi(\lambda_{bottom})\lambda_{bottom} - \phi(\lambda_{top})\lambda_{top} \\ & + (1 - Q(\lambda_{bottom}))\lambda_{bottom}^2 + Q(\lambda_{top})\lambda_{top}^2) \end{aligned} \quad (19)$$

Effective electrical signal to noise ratio per bit ($SNR/bit_{electrical}$) can be represented as a function of undistorted $SNR/bit_{electrical}$, $\gamma_{b(e)} = \frac{E_{b(e)}}{N_0}$ as follows,

$$SNR = \frac{K^2}{\frac{G_B \sigma_{clip}^2}{\sigma_y^2} + \frac{G_B \gamma_{b(e)}^{-1}}{G_{DC}}} \quad (20)$$

where $G_B = 0.5$ for ACO and $(N - 2)/N$ for DCO and G_{DC} is the gain factor denoting attenuation of signal due to DC component, expressed as,

$$G_{DC,DCO} = \frac{\sigma_y^2}{\sigma_y^2 + \beta_{DC}^2} \quad (21)$$

$$G_{DC,ACO} = \frac{\sqrt{2\pi}\sigma_y^2}{\sqrt{2\pi}\sigma_y^2 + 2\sqrt{2\pi}\beta_{DC}^2 + 4\sigma\beta_{DC}} \quad (22)$$

β_{DC} is the DC bias.

For M-QAM, BER performance in additive white Gaussian noise (AWGN) [9] is given as,

$$BER = \frac{(4\sqrt{M} - 1)}{\sqrt{M}\log_2(M)} Q\left(\sqrt{\frac{3\log_2(M)}{M-1} SNR}\right) + \frac{(4\sqrt{M} - 2)}{\sqrt{M}\log_2(M)} Q\left(3\sqrt{\frac{3\log_2(M)}{M-1} SNR}\right) \quad (23)$$

PAPR is defined as,

$$PAPR = \frac{\max(|y_m^2|)}{E[y_m^2]} \quad (24)$$

OFDM signal generally suffers from the high PAPR issue. PAPR is usually shown with the help of complementary cumulative distribution function (CCDF) plot.

Following the procedures elaborated in [10], PAPR CCDF of the clipped DCO-OFDM can be written as,

$$F_{DCO}(y) = \begin{cases} 1 - [2\phi(\sqrt{s * y}) - 1]^N & ; 0 \leq y < \theta_{min} \\ 1 - [\phi(\sqrt{s * y})]^N & ; \theta_{min} < y < \theta_{max} \\ 0 & ; y \geq \theta_{max} \end{cases} \quad (25)$$

PAPR CCDF of the clipped ACO-OFDM signal is given as,

$$F_{ACO}(y) = \begin{cases} 1 - [2\phi(\sqrt{s * y}) - 1]^{N/2} & ; 0 \leq y < \theta_{min} \\ 0 & ; y > \theta_{max} \end{cases} \quad (26)$$

where,

$$\theta_{min} = \frac{\lambda_{bottom}^2}{s} \quad \text{and} \quad \theta_{max} = \frac{\lambda_{top}^2}{s} \quad (27)$$

$$s = (\lambda_{bottom}^2 - 1)\phi(\lambda_{bottom}) - (\lambda_{top}^2 - 1)\phi(\lambda_{top}) + \lambda_{top}^2 + \lambda_{bottom}g(\lambda_{bottom}) - \lambda_{bottom}g(\lambda_{bottom}) \quad (28)$$

; $g(\cdot)$ is the PDF of the standard normal function.

III. RESULTS AND DISCUSSIONS

The system performance of IM/DD optical systems is simulated using MATLAB software. Introduced uncorrelated noise due to dual-side signal clipping affects the BER and PAPR performance of the system. A wise selection of these clipping levels needs to be done. Fig. (2) and Fig. (3) shows the effect of λ_{bottom} and λ_{top} on the K and σ_{clip}^2 in DCO-OFDM and ACO-OFDM respectively. σ_{clip}^2 tends to zero for scenarios with minimal signal clipping. Since σ_{clip}^2 and K are dependent on each other according to eqn.(16,18), as K tends to zero, σ_{clip}^2 also tends to zero. This can be verified from Fig. (2) and Fig. (3).

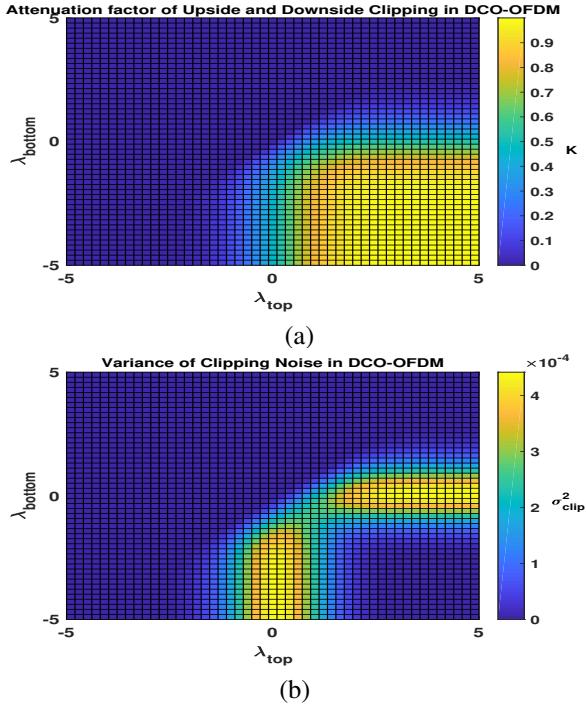


Fig. 2: Effect of normalized bottom (λ_{bottom}) and top (λ_{top}) clipping levels in DCO-OFDM on (a) K (b) σ_{clip}^2 .

Large σ_{clip}^2 and small K implies large symbol distortion. Thus, an inference can be derived from Fig. 2(a) and Fig. 2(b) that the symmetric clipping minimizes the distortion of the symbol in DCO-OFDM. While Fig. 3(a) and 3(b) suggests that the downside clipping distorts the symbol to a larger extent than upside clipping in ACO-OFDM. Thus, upside clipping is preferable in case of ACO-OFDM.

From the eqn. (15-19) discussed above, it can be seen that K and σ_{clip}^2 depends only on the λ_{top} and λ_{bottom} , independent of the size of IFFT (F) and the modulation order (M). Thus, K and σ_{clip}^2 remains constant for any selected modulation order, keeping the normalized top and bottom clipping levels same.

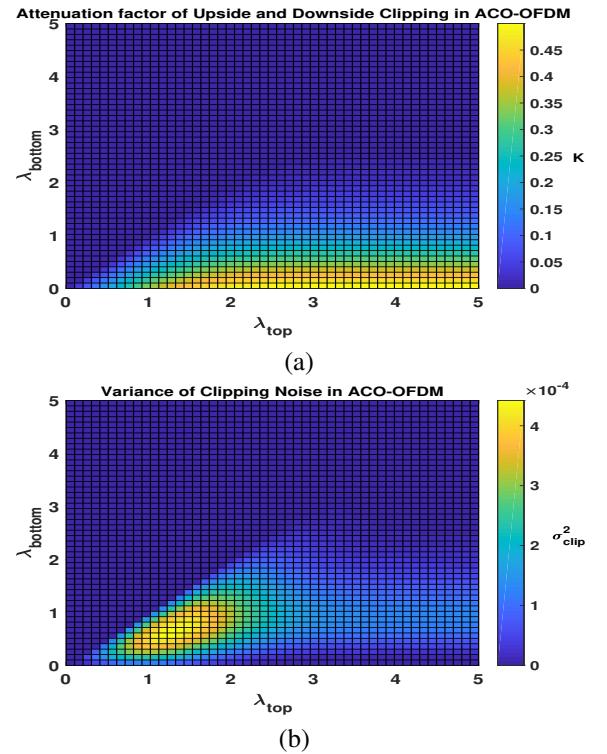


Fig. 3: Effect of normalized bottom (λ_{bottom}) and top (λ_{top}) clipping levels in ACO-OFDM on (a) K (b) σ_{clip}^2 .

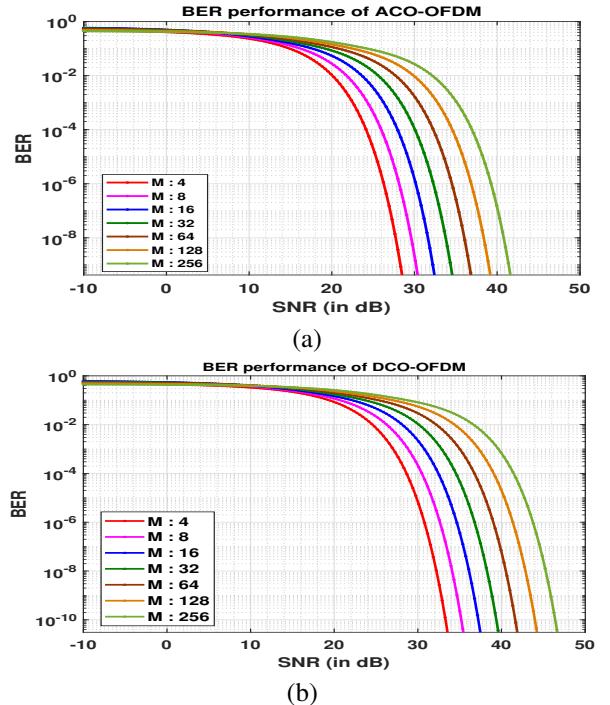


Fig. 4: BER performance on various modulation index (M) for (a) ACO-OFDM (b) DCO-OFDM.

BER of the system depends on the modulation order and as the modulation order increases, the exposure of the constellation points to Gaussian noise increases due to the contraction

of the decision regions. Thus, for a particular chosen values of the clipping levels in the transmitter dynamic range, additive noise due to dual-sided clipping distortion affects the BER performance. Mathematically, the dependency of the BER on the clipping levels can be seen from eqns. (20-23). It can be concluded from the Fig. 4(a) and 4(b) that as M increases, BER of the ACO-OFDM and DCO-OFDM system deteriorates (i.e. higher SNR is needed for achieving a particular value of BER), respectively.

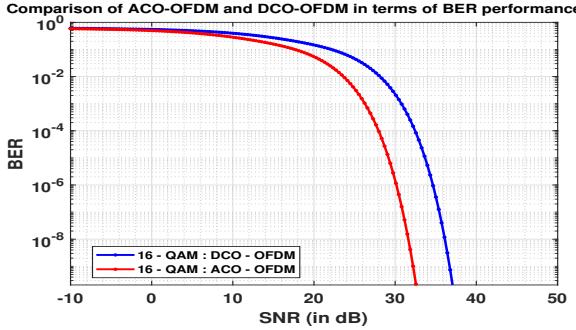


Fig. 5: BER comparison of DCO-OFDM and ACO-OFDM for VLC.

Comparing the performance of the DCO-OFDM and ACO-OFDM configuration, Fig. (5) shows that ACO-OFDM outperforms DCO-OFDM in higher SNR region. In order to achieve a BER of 1×10^{-3} , ACO-OFDM shows an advantage of about 4.5 dB as compared to DCO-OFDM.

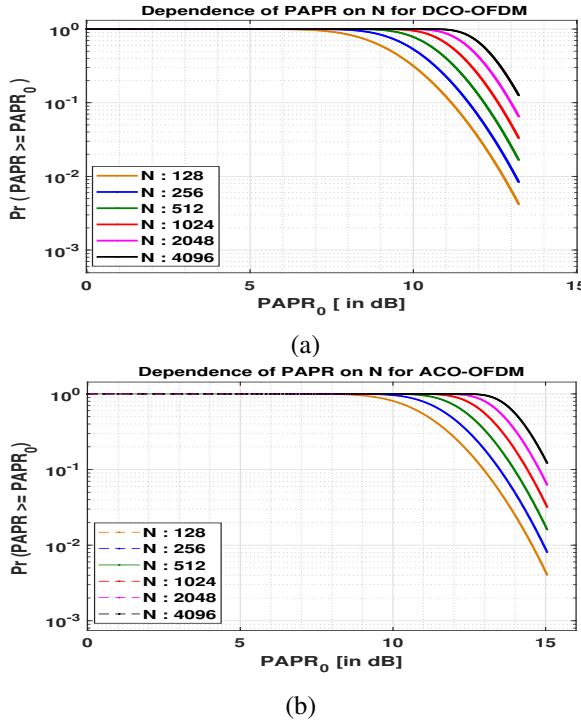


Fig. 6: Dependence of PAPR on number of subcarriers (N) for (a) DCO-OFDM (b) ACO-OFDM.

The PAPR of the ACO-OFDM and DCO-OFDM system

is a very critical parameter. In order to study the parameters affecting PAPR, simulations are done to determine the effect of number of subcarriers (N) and selection clipping levels. Large N means narrow bandwidth of each subcarrier implying longer symbol period, where the impact of non-linearity grows. This leads to the ISI. But the effect of ISI is not so severe in the case of VLC systems because of the slow fall-off of the VLC channel. Fig. 6(a) and 6(b) shows the dependence of PAPR on N for DCO-OFDM and ACO-OFDM systems. It can be seen that the PAPR increases with the increase in N . Since we know that, large N requires higher computational complexity. So, lower N is preferable as far as PAPR performance is concerned.

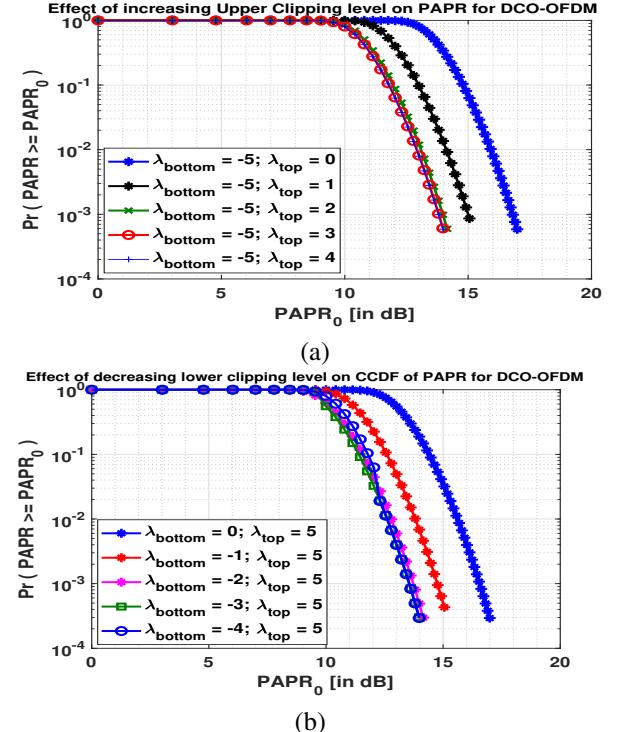


Fig. 7: Effect on PAPR for DCO-OFDM of (a) increasing upper clipping level (b) decreasing lower clipping level.

A wise selection of the normalized clipping level is advised since it also affects the PAPR performance. Fig. 7(a) and 7(b) shows the effect of increasing the λ_{top} and decreasing the λ_{bottom} for DCO-OFDM. $\theta_{min.}$ and $\theta_{max.}$ are the dividing points in the piecewise PAPR CCDF of the DCO-OFDM. In case of asymmetric clipping, it can be seen that the single-sided clippings increases the $PAPR_{max}$ values, i.e. as the extent of clipping of signal from one side (bottom or top) increases, the maximum value of PAPR also increases.

But in case of symmetric clipping, PAPR CCDF has only one dividing point (i.e. $\theta_{max.}$) and larger the extent of clipping, lesser the value of the maximum value of PAPR as can be seen in Fig. (8). In other words, symmetric clipping reduces the PAPR. This can also be verified from the Fig. 3(a) and 3(b), where symmetric clipping reduces the large symbol distortion. While for ACO-OFDM, as the extent of clipping from above

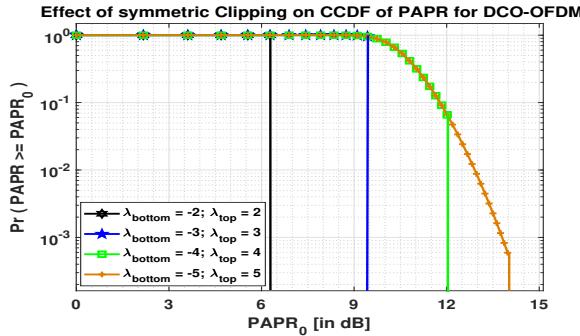


Fig. 8: PAPR CCDF for DCO-OFDM with symmetrical clipping

increases, maximum PAPR reduces as can be seen in Fig. (9). This is in accordance with the result derived from the Fig. 2(a) and 2(b) showing that for ACO-OFDM, the upside clipping is preferred.

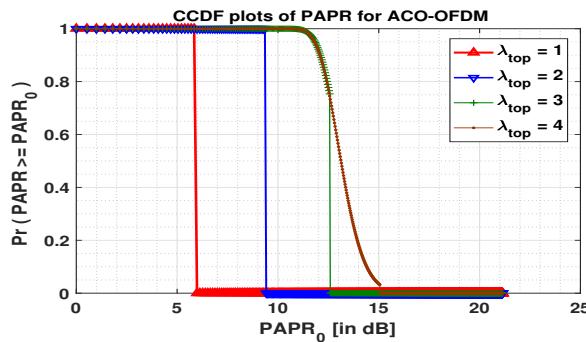


Fig. 9: Variation of PAPR CCDF for ACO-OFDM with upside clipping.

Comparing the PAPR of the DCO-OFDM and ACO-OFDM, Fig. (10) shows that for a CCDF of 0.1, ACO-OFDM outperforms DCO-OFDM by about 2 dB. The reason behind this is the addition of significant amount of DC bias to make signal non-negative in DCO-OFDM, which increases the peak power of the sub-carriers whereas in ACO-OFDM, signal is clipped at zero level eradicating the use of bias for this purpose. Thus, resulting in reduced PAPR as compared to DCO-OFDM.

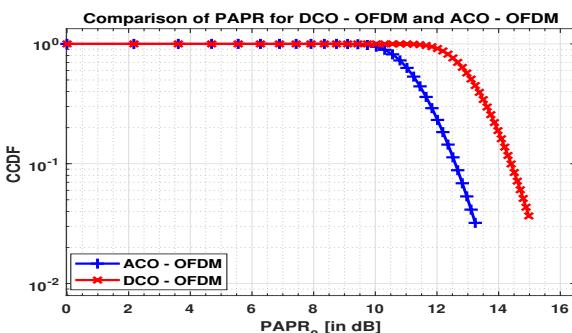


Fig. 10: PAPR comparison of DCO-OFDM and ACO-OFDM.

IV. CONCLUSION

In this paper, analysis of BER and PAPR performance for DCO-OFDM and ACO-OFDM in IM/DD optical systems is done. Analysis of wise selection of clipping levels for reducing the PAPR is done. DCO-OFDM is less efficient in terms of power efficiency but has higher spectral efficiency as compared to ACO-OFDM. The dependence of BER on the order of modulation is analyzed. It was found that as the modulation order increases, BER performance deteriorates. While comparing these two schemes, ACO-OFDM outperforms DCO-OFDM by around 4.5 dB to achieve a BER of 10^{-3} . The effect of the dual-sided clipping on the signal is analyzed in terms of the clipping noise variance and attenuation factor. It was found that for ACO-OFDM, upside clipping is preferred while in case of DCO-OFDM, symmetric clipping is preferred in order to reduce the symbol distortion. This setup of clipping levels reduces the maximum PAPR value for these systems. The number of subcarriers affects the PAPR performance. It was found that as the number of subcarrier increases, PAPR increases. Also, the large number of subcarriers require higher computational complexity. So, lower number of subcarriers are preferred as far as PAPR performance is preferred. On comparing PAPR, ACO-OFDM outperforms by about 2 dB for a PAPR CCDF of 10^{-1} . The results of the paper can be used for evaluation of the proper DC biasing and signal spacing for the specified dynamic range of LED resulting in optimized electrical SNR and BER.

ACKNOWLEDGMENT

This work is supported by Department of Telecommunications (DoT), project ‘Indigenous 5G Test Bed’, No. 4-23/5G-Test Bed/2017-NT.

REFERENCES

- [1] D. C. O'Brien, L. Zeng, H. Le-Minh, G. Faulkner, J. W. Walewski and S. Randel, “Visible light communications: Challenges and possibilities,” *IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications*, Cannes, France, pp. 1-5, Sept. 2008.
- [2] IEEE Standards Association, [Available online]: <http://standards.ieee.org/develop/project/802.15.7.html>.
- [3] D. Karunatilaka, F. Zafar, V. Kalavally, and R. Parthiban, “LED Based Indoor Visible Light Communications: State of the Art,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 3, pp. 1649-1678, 2015.
- [4] J.M. Kahn and J. R. Barry, “Wireless infrared communications,” *Proc. IEEE*, vol. 85, no. 2, pp. 265-298, Feb. 1997.
- [5] S. Dimitrov, S. Sinanovic, and H. Haas, “Signal shaping and modulation for optical wireless communication,” *J. Lightw. Technol.*, vol. 30, no. 9, pp. 1319-1328, May 2012.
- [6] A. Yang, Y. Wu, M. Kavehrad, and G. Ni, “Grouped modulation scheme for led array module in a visible light communication system,” *IEEE Wireless Communications*, vol. 22, no. 2, pp. 24-28, April 2015.
- [7] L. Mao, C. Li, H. Li, X. Chenb, X. Maob and H. Chenb, “A mixed interval multi-pulse position modulation scheme for real-time visible light communication system,” *Optics Communications*, vol. 402, pp. 330-335, Nov. 2017.
- [8] Z. Wang, T. Mao and Q. Wang, “Optical OFDM for visible light communications,” *13th International Wireless Communications and Mobile Computing Conference (IWCMC)*, Valencia, pp. 1190-1194, 2017.
- [9] S. Dimitrov, S. Sinanovic, and H. Haas , “Clipping Noise in OFDM-Based Optical Wireless Communication Systems,” *IEEE Transactions on Communications*, vol. 60, no. 4, pp. 1072-1081, April 2012.
- [10] J. Wang, Y. Xu, X. Ling, R. Zhang, Z. Ding, and C. Zhao , “PAPR analysis for OFDM visible light communication,” *Optics Express*, vol. 24, no. 24, pp. 27457-27474, 2016.

Modelling and short term forecasting of flash floods in an urban environment

Suraj Ogale

Dhirubhai Ambani Institute
of Communication and
Technology,Gandhinagar
Gujarat 382007

Email: ogalesuraj@gmail.com

Sanjay Srivastava

Dhirubhai Ambani Institute
of Communication and
Technology,Gandhinagar
Gujarat 382007

Email: sanjay_srivastava@daiict.ac.in

Abstract—Rapid urbanization, climate change, and extreme rainfall have resulted in a growing number of cases of urban flash floods. It is important to predict the occurrence of a flood so that the aftermath of it can be minimized. As the name suggests, an urban flash flood occurs in an urban area in a very short span of time. To reduce the impact of these events, short-term forecasting or nowcasting is used for prediction of the very near future incident. In orthodox methods of flood forecasting, current weather conditions are examined using conventional methods such as the use of radar, satellite imaging and calculations involving complicated mathematical equations. However, recent developments in Information and Communication Technology (ICT) and Machine Learning (ML) has helped us to study this hydrological problem from a different perspective. The aim of this paper is to design a theoretical model considering the parameters causing the urban flash flood and predict the event beforehand. To test the soundness model, data syntheses is performed and the results are checked using the artificial neural network.

I. INTRODUCTION

Flood is one of the biggest natural disaster causing many lives as well as damages. Different types of floods like river flood, urban flood, coastal flood, and flash flood have been observed over the years. A flash flood is a direct response to a rainfall having very high intensity in small time. This kind of flood is seen typically in urban areas where the underlying ground cannot cope, or drain excess water away fast enough via the sewage system and draining canals in a short amount of time. In recent years, we have seen the impact of floods in cities such as Mumbai(2005), Chennai(2015), Ahmedabad(2017). Poor urban planning, inaccurate and delayed forecasting and inadequate flood mitigation system are the main reasons behind it. The conventional strategies of flood forecasting are expensive and highly complex. Weather and rainfall forecasting is a major task behind the prediction of a flood. Weather forecasting involves simulations based on physics and differential equations. The rainfall forecast is done using radars and satellite imaging. A Doppler weather radar is used to locate the precipitation and detect the motion of rain droplets. Dedicated weather satellite provides images using which information about rainfall can be deduced.

Our approach for short-term flash flood prediction in urban

areas is to establish a theoretical model incorporating the factors influencing flood and use the power of machine learning techniques to estimate flood ahead of time. The organization of this paper is as follows. In section II, we have discussed existing techniques used in the field of flood forecasting from a machine learning perspective. In section III, we have described the problem undertaken and methodology behind the proposed model. In section IV, the Implementation strategy is described. In section V, results obtained by the neural network are discussed.

II. RELATED WORK

Much of the research work has been carried out in river flood forecasting whereas urban flash flood has been explored a little. Wireless Sensor Networks for Flash-Flood Alerting by Castillo-Effer et al.[1] discusses the use of wireless sensor networks (WSN) in the Andean region of Venezuela. WSN is used for monitoring the environment and tracking the disaster while it evolves. Liong, S.Y. et al.[2] studied the problem of flood stage forecasting using support vector machines (SVM) in the region of Dhaka, Bangladesh. In this paper, data from 8 water stations are taken into consideration and the next water level has been predicted using SVM. Results obtained from both SVM(Support Vector Machine) and ANN(Artificial Neural Network) are compared. Mousa Mustafa et al.[3] put forth the use of non-contact sensors in the study of Flash Flood Detection in Urban Cities Using Ultrasonic and Infrared Sensors. Lekkas et al.[4] studied the effect on the performance of prediction by evaluating different types of ANN. The case study presented shows that ANN can be used for flood forecasting. Because of the property of universal approximation, neural networks are used widely in the field of hydrology. Flash Floods Forecasting without Rainfalls Forecasts by Recurrent Neural Networks by Artigue g. et al.[5] discusses the use of a neural network to predict the discharge rate. They have taken a case study in France at mialet basin that consists 3 locations. The Problem discussed by the paper is to predict the discharge rate $q(k + f)$ i.e. outflow at time $k + f$, where k is current time and f is a future window for ahead prediction. A sliding window w is chosen with respect to distance from mialet, rain intensity and impact of rain. Over this past window

w , rainfall values are collected. But only relying on this input is not sufficient, since current discharge will have an impact on the next discharge. Thus, the estimated discharge at time $k-1$, $q(k-1)$ is also taken as input variable. A Nash criteria[6] is used to quantify the quality of forecasts. The optimal desired value of nash criteria should be closer to 1. Authors have obtained the value of nash criteria in the range of 0.7 to 1. Kei Hiroi and Nobuo Kawaguchi[7] present a system which performs real-time river level monitoring and flood prediction for an urban flow particularly caused by localized heavy rain(LHR). Authors have developed a compact sensor design consisting of three parts: an infrared camera, a communication interface and an image processing server. Using this designed sensor, the water level is detected. Along with two patterns of rainfall: Last 10 minutes rainfall and rainfall of past fixed window, monitored water level is fed to a server which accurately predicts water level after 5 minutes. To do prediction of flood, long short-term memory architecture of LR is used for prediction.

III. PROBLEM STATEMENT AND PROPOSED APPROACH

In the urban cities, sub-pass or a low-lying area is most vulnerable to waterlogging. On this flood-prone locations, water gets accumulated in a short period of time. Relative elevation, surface runoff and insufficient passage of water to drainage are key points in the development of waterlogging. Thus, flood forecasting at these places is essential. In this work, we have proposed to do modelling of urban city and estimate the stage of a flash flood in near future particularly at places which are at the risk of flooding. We devise to create cells by dividing an area into many regions having alike properties. Factors affecting flood and methods for gathering them are discussed. There exists a pattern between all the considered parameters and flood. To learn this relationship, we need a prediction technique. Artificial neural networks are a family of models inspired by biological neural networks. It has the ability to learn the complex relationship between input and output. It can find any non-linear relation between subjected features and thus chosen over other methods of prediction.

This system is proposed with a few assumptions:

- 1) Rainfall is distributed over a cell.
- 2) Soil index is constant between storms and is determined solely by surface condition and type.
- 3) Timescale of the change in the value of the clogging factor is large. It is gradual changing property.
- 4) Methods used to measure each parameter are chosen to minimize any kind of error. These methods are not compulsory to choose and alternative methods can be explored.

A. Hydrological Model

There exists data-driven deterministic models[8] which uses precipitation modelling and flood routing. These models' performance and efficiency depends upon catchment and requires high-quality data and modelling expertise to produce accurate results. The data-driven models are black-box methods as they

depend on the statistical relationships between parameters. Our data-driven model is as follows:

1) *Cell Creation:* We are dividing an urban area into few cells having alike properties. These cells are created based on the below characteristics:

- (a) Size and Elevation of the catchment(cell)
- (b) Soil and/or surface characteristics
- (c) Drainage outlets

A cell is the subpart of a city which has similar properties over its area. The cells are static in nature and can be made of any arbitrary shape. Structure of cell is changed if a variation in its property or any physical change is found in a cell. Few Examples of cells are:

- (i) A section of road, where it has the same elevation, surface property and drain out.
- (ii) A cluster of buildings sitting on an equal slope and surface with common drainage.
- (iii) A road-bridge since it has different elevation and does not absorb any water.

With respect to the division of cells, we create a directed graph where each cell is denoted by a vertex. A directed edge between two vertices (say x and y) is added only if water from one cell (x) goes to another cell (y) as surface runoff. The direction of an edge depends upon the flow from one cell to other.

2) Features used to create cells:

- (a) Size and elevation of catchment

A catchment is an important factor behind flooding in which its size and elevation change impacts the flood. Size will decide the amount of water going to drainage or going to another part of the city. Also, the shape and elevation affect the speed and time water takes to reach drainage and other adjacent catchments. Time of concentration is the time taken for runoff water from the furthest point of the catchment to reach the point of interest say sewer inlet. The size, shape, and elevation affect this time of concentration.

- (b) Soil and/or surface characteristics

The amount of water absorbed depends heavily on the kind of soil ground has. The infiltration capacity varies from soil to soil. For a particular soil type, this capacity also changes over time. Based on the soil absorption index, water infiltration varies.

- (c) Drainage outlets

Each city has its drainage network laid out. The amount of water that drains out from a cell depends heavily on this drainage network. Also, the amount of clogging affects the rate at which the drainage pipes can dispose of water.

3) Inputs to the system:

- (a) Rainfall:

It is the primary reason behind the flooding situation. The intensity of rain largely impacts the gravity of flood. For high-intensity rain, the time needed for peak flow is less and vice versa. If the intensity is constant than the

duration of rainfall determines the peak flow and time period of surface runoff.

(b) Surface runoff from adjacent cells:

With respect to the relative elevation and adjacency of cells, water will definitely move from one cell to other. The adjacency and direction of runoff from one cell to another depends upon the structure of cells.

(c) Drainage capacity and clogging factor:

Each drainage pipe and thus the entire network has some threshold rate at which it can operate faultlessly. Above this threshold, the network may behave erratically. The drainage network can be difficult to keep clean and has some clogging over the time. This clogging factor changes gradually over time.

(d) Water sent to drainage:

from every cell having a drainage outlet, water will be taken out via drainage. The amount of water drained out from the cell depends on the drainage capacity as well as the clogging factor.

(e) Current water level in the cell:

All the effects of rainfall will be eventually on the height of water or water level in the cell. This impact will either increase or decrease the current cell water level.

(f) Rain forecast:

The above five factors talk about past events. Along with this, future rainfall will have an influence on a flood. According to the accuracy of short-term radar-rainfall forecasts [9] by Bellon and Austin, radar-based accumulations have an inherent error of the order of 25% in which 0.5 hr forecasts have an error of 50% and 3 hr forecasts have an error of 60%. This rain forecast is made available using Radar. Thus rain forecast is included as an input to the system.

4) Model creation and output formulation: Our model is divided into 2 parts:

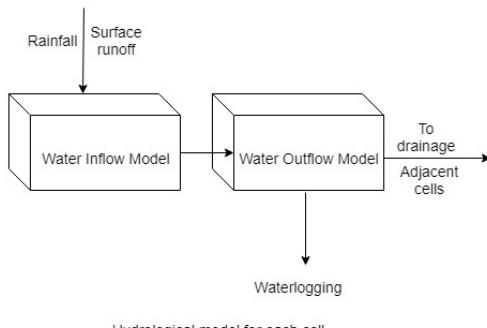


Fig. 1: Block Diagram of a system

(a) Water inflow model

This is the first component of the model where water coming to the cell is examined. The inflow of water constitutes rainfall and surface runoff from adjacent cells.

(b) Water outflow and waterlogging model

Water going via drainage and to adjacent cell constitutes net outflow from the cell. Since each drainage has fixed capacity and if the rate of water sent to drainage is more than this capacity, backwater current will get created. It will get added to the amount of water in the cell. Methods for measuring these features are discussed in section IV.

To formulate the factors into the equation, consider two cells(A and B) adjacent to each other such that A is relatively lower to B with respect to elevation. Cell A is a candidate for waterlogging. Figure 2 shows the relative structure of these cells. A past sliding window w is fixed over which all the

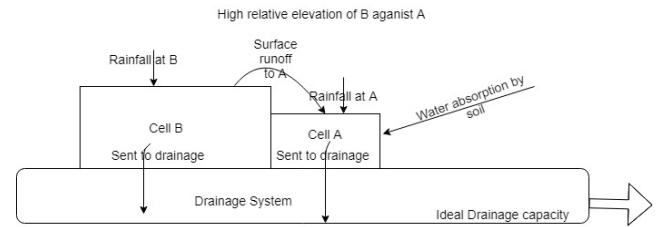


Fig. 2: Example of two adjacent cells

data is collected. Thus, the water level in cell A at time $t + 1$ is given as:

$$h_A(t + 1) = h_A(t) + f(x) \quad (1)$$

where

x is the residual water left in a cell.

$h_A(t + 1)$ and $h_A(t)$ are height of water over surface in cell A at time $t + 1$ and t respectively.

Function $f(x)$ has to be learned since it has a behaviour related to subjected cell, its adjacent cells and their respective properties.

Now, x can be written as:

$$x = [\text{waterin}] - [\text{waterout}] + [\text{surplus water against drainage}]$$

which is further simplified as:

$$x = [r_A + d_B] - [a_A + c_A] + [c_A - DC(1 - \alpha_{clogging})] \quad (2)$$

where,

r_A : Rainfall in cell A

d_B : Discharge/Surface runoff from cell B.

a_A : Water absorption by soil

c_A : Water sent to drainage

$\alpha_{clogging}$: A constant measuring percentage of clogging.

DC : Ideal Drainage capacity

$DC(1 - \alpha_{clogging})$: Actual drainage capacity

All the variables except $\alpha_{clogging}$ (numeric 0 to 1) and height (in cm) are rates in unit volume per time.

The term $[c_A - DC(1 - \alpha_{clogging})]$ gives us the flow of the water that runs against the drainage. It is possible only when

the rate of water sent to drainage is more than drainage can handle. This term will be taken into account only when the $c_A > DC(1 - \alpha_{clogging})$. In other cases, it will be ignored totally.

Thus when $c_A > DC(1 - \alpha_{clogging})$, rewriting (1):

$$h_A(t+1) = h_A(t) + f[r_A + d_B - a_A - DC(1 - \alpha_{clogging})] \quad (3)$$

When no reverse flow of water is present, we will have:

$$h_A(t+1) = h_A(t) + f[r_A + d_B - a_A - c_A] \quad (4)$$

IV. IMPLEMENTATION STRATEGY

To practically implement this method, the cells are created using the features said earlier. Elevation of any location can be made available via altitude or elevation maps that are obtained by satellites. This elevation can be either in form of height above sea level or relative elevation against neighbouring cells. Soil index is measured for each cell and drainage outlet is checked along with their capacity. To measure the water sent to drainage and water coming from the adjacent cells, open flow methods are used such as Weir and Flume having the hydraulic structure as well as methods like Dye Testing and Manning's theoretical equation[10].

To check the validity of the proposed model, it has to be tested on real data and the performance of the model must be examined. Since no actual data is present, a synthesis is done where we let the pattern of water level at time $t+1$ to generate using the values of net residue water in the cell and rainfall forecast. The 6-dimensional input to our system is listed as rainfall, water from adjacent cells, water absorbed, a surplus of water when sent to drainage against drainage capacity, current water level in the cell, and rain forecast whereas the output is the stage of a flood. The input is considered over the sliding window of $w = 2$ hour and the forecast is done ahead at time $t+1$ hours which we call as future lookahead time. The stages of a flood are decided by the water level in the cell. The water level is divided into 4 bands and these 4 bands depict the stages of a flood from low intensity to high. One-hot encoding is used to convert the numeric value to a class. The 4 stages with respect to severity of flood are as follows:

Severity	Description
0	No flood
1	Minor
2	Moderate
3	Major

TABLE I: Flood stages

A. Dataset Generation

Multiple scenarios were constructed in which dataset is generated. Combinations of continuous values, samples from normal distribution, constant random values as well as samples from functions(such as quadratic, trigonometric) were studied. One such synthesis for the features is given in Table II.

Data for each input feature is generated independently with randomness. Rainfall, rain forecast and the current water level

Feature	Generation
Rainfall and rain forecast	Samples from a normal distribution
Water absorption	Fixed part of rainfall
Water to drainage and Clogging factor	Samples from a uniform distribution
Drainage capacity	Random fixed number

TABLE II: Feature generation

is generated from the normal distribution. Since the normal distribution shows up frequently in nature, we have considered the data in this pattern. This dataset contains 20000 samples and is created such that each flood stage contains evenly proportionate tuples.

B. Scenarios constructed for simulations

Data synthesis is done for two scenarios. The first scenario contains only two adjacent cells and we are interested in waterlogging at cell A whereas the second scenario has seven cells and we are looking for possible flooding in cell E.

1) Model run 1 for Two cells scenario:

According to Figure 2, cells A and B are created. Data Generation of input is done according to Table II.

The output is produced using below equation:

$$h(t+1) = h(t)(1 + \delta(t)) \quad (5)$$

where $\delta(t) = x(t) + RF(t+1)$

x is water residue and $RF(t+1)$ is rain forecast for time t to $t+1$ in the range of 0 to 1.

2) Model run 2 for Multi-Cell scenario: For the testing of a complex scenario, we have imagined a fictional area in the urban city. In this area, we have assumed that a low-lying area a sub-pass exists which is vulnerable to flood. This area constitutes one cell and neighbouring cells of roads and buildings are visualized. Thus seven cells are created which reciprocates a real-world example. The sub-pass is surrounded by a cluster of buildings, roads and ground so that possible variation with respect to the features that divides cells is covered. Each cell has its own characteristic for size, elevation, drainage outlets, and soil type. Each cell has its drainage outlet whereas the elevation is in terms of low, medium and high. Cell A, B, and C are roads whereas D and G are clusters of buildings. Cell F is ground and cell E is sub-pass which is the flood-prone area. Figure 3 shows the graph

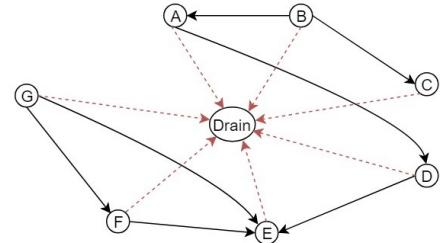


Fig. 3: Graph for multiple cells

for seven cells where each cell is represented as vertex and the edges between two vertices represent the water

flowing from one cell to another. The dashed edges show the water flowing to drainage. In this experiment, we are interested in the flood stage at cell E. For each cell, the calculation is done independently and beforehand such that it leads to water coming from the adjacent cells of E to cell E.

For both scenarios, the dataset is fed to the neural network containing 5 layers including the input layer which has 6 neurons and the remaining layers have nos of neurons as 60,30,15 and 4 respectively where the last layer is output layer.

The output of hidden layer neurons and the last layer neuron will be a linear function in the absence of activation function. So, to learn more complex and non-linear functions we have to throw some kind of nonlinearity in the network. This non-linearity is introduced in the form of activation function in the network. In hidden layers, the most popular activation function, ReLU is used. ReLU stand for "Rectified Linear Unit" and is represented as $f(x) = \max(0, x)$. The softmax activation function is used in the output layer as the output of the neural network is classes of the flood. Adam optimizer is used in the neural network since it is computationally efficient. In this, a learning rate is maintained for each network weight (parameter) and separately adapted as learning unfolds. Adam is a popular algorithm in the field of deep learning because it achieves good results fast. In our problem, for the loss function, we are using cross entropy loss for multi-class classification. Multiple structures of neural network were inspected before finalizing this structure. The neural network is trained by doing 100 epochs. Other hyperparameters are set to their standard practised values. Below is the list of hyperparameters whose values are set before the start of a learning process.

Parameter	Value
Loss function	Categorical crossentropy
Optimizer	Adam
Epochs	100
Validation split	0.33
batch size	100

TABLE III: Hyperparameters in neural network

V. RESULTS AND ANALYSIS

- 1) According to the Implementation Strategy, training of neural network is done for both two cells as well as multiple cells scenario. To create the neural network, Keras and tensorflow libraries are used. The dataset contains 20000 samples and is divided into three sections viz training, validation and testing datasets.
- 2) For both scenarios, accuracy is measured for each dataset. Here, accuracy measures the percentage of how many classes are correctly classified over the entire dataset. For two cells, accuracy is found as:
- 3) For multi-cells scenario, 3 different iterations are done each having subtle tweaks in the way data of water

Dataset	Accuracy
Training	97%
Validation	95%
Test	94%

TABLE IV: Results for two cells scenario

residue and rain forecast is generated and the results are obtained. The iterations are as follows:

- Iteration 1: Water residue normalized from -1 to 1 and rain forecast is binary(0: No rain /1: Rain).
- Iteration 2: Water residue normalized from -1 to 1 and rain forecast is continuous valued from 0 to 1.
- Iteration 3: Water residue and rain forecast both continuous valued.

The results for each iteration can be seen in Table V.

Dataset accuracy	Iteration 1	Iteration 2	Iteration 3
Training	96%	97%	94%
Validation	93%	96%	94%
Test	93%	93%	90%

TABLE V: Results for three iterations

A. Insights deduced from the simulations and results:

- 1) The numerical aspect is just one facet for checking the obtained results but the trend or pattern from the results also serves as a great indicator. Few other simulations along with the above mentioned are done in order to see the insights.
- 2) Flash floods normally occur in the very short amount of time, thus we need to do short-term forecasting of these events. The short-term forecasting includes prediction of the event ahead in 0-6 hours. So in our model, prediction is done for time $t + 1, t + 2, t + 4$ and $t + 6$ hours ahead where t is current time. For these types of prediction methods, recall is also an important parameter. A recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances.
- 3) Of the all four flood stages that we have defined, most accurate results were seen in stage zero (No flood) and stage one (Minor) whereas little error was seen in the accuracy of stage two (Moderate) and three (Major). The recall for stage 2 and 3 was lower than the recall seen for stage 0 and 1.
- 4) The accuracy of the four scenarios where we changed future lookahead time varies accordingly. As we increase the future lookahead time, accuracy decreases. For $t + 1$ scenario, accuracy seen was maximum among all four cases. As we increase the lookahead time, the accuracy of flood prediction drops considerably.
- 5) The reason behind this drop in efficiency of a model is twofold:
 - (a) As the future lookahead time increases, the dependence of the input variables against the output of flood stages weakens out.

- (b) The rainfall forecast also introduces more error since the near term rain forecast is much more efficient and less error-prone against the long-term rain forecast.

VI. CONCLUSION

There were limited attempts in development of flash flood modelling system but not in the area of urban flash floods. This study is aimed at the modelling and prediction of urban flash floods. Our first step was to sketch a model for an urban area in which short-term forecasting of a flood can be done. The cells are created keeping in mind the properties that determine the urban flash flood. A template of two cells is constructed on which simple equations have been formulated. Also, a complex scenario including multiple cells is imagined where a possible real-world scenario is undertaken. For these cases, data simulation and the sample results are examined. Insights from the simulations and subsequent results put forth important aspects of the short-term flood forecasting. It tells us that the lookahead time for flash flood prediction should be kept in the optimal period.

Our future endeavour and the second step is to develop a prototype so that it can be applied in actual condition. By collecting true data, the repository would be built. On this dataset, the result would be seen using a neural network. Better techniques for modelling, measuring the parameters especially using sensors node would be developed.

REFERENCES

- [1] M. Castillo-Effer, D. H. Quintela, W. Moreno, R. Jordan, and W. Westhoff, "Wireless sensor networks for flash-flood alerting," in *Devices, Circuits and Systems, 2004. Proceedings of the Fifth IEEE International Caracas Conference on*, vol. 1. IEEE, 2004, pp. 142–146.
- [2] S.-Y. Lioung and C. Sivapragasam, "Flood stage forecasting with support vector machines," *JAWRA Journal of the American Water Resources Association*, vol. 38, no. 1, pp. 173–186, 2002.
- [3] M. Mousa, X. Zhang, and C. Claudel, "Flash flood detection in urban cities using ultrasonic and infrared sensors," *IEEE Sensors Journal*, vol. 16, no. 19, pp. 7204–7216, 2016.
- [4] D. Lekkas, C. Onof, M. Lee, and E. Baltas, "Application of artificial neural networks for flood forecasting," *Global Nest Journal*, vol. 6, no. 3, pp. 205–211, 2004.
- [5] G. Artigue, A. Johannet, V. Borrell, and S. Pistre, "Flash floods forecasting without rainfalls forecasts by recurrent neural networks. case study on the mialet basin (southern france)," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, Oct 2011, pp. 303–310.
- [6] J. Nash and J. Sutcliffe, "River flow forecasting through conceptual models part i to a discussion of principles," *Journal of Hydrology*, vol. 10, no. 3, pp. 282–290, 1970. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/0022169470902556>
- [7] K. Hiroi and N. Kawaguchi, "Floodeye: Real-time flash flood prediction system for urban complex water flow," in *SENSORS, 2016 IEEE*. IEEE, 2016, pp. 1–3.
- [8] S. K. Jain, P. Mani, S. K. Jain, P. Prakash, V. P. Singh, D. Tullus, S. Kumar, S. Agarwal, and A. Dimri, "A brief review of flood forecasting techniques and their applications," *International Journal of River Basin Management*, pp. 1–16, 2018.
- [9] A. Bellon and G. Austin, "The accuracy of short-term radar rainfall forecasts," *Journal of hydrology*, vol. 70, no. 1-4, pp. 35–49, 1984.
- [10] R. Manning, "On the flow of water in open channels and pipes," *Transactions of the Institution of Civil Engineers of Ireland*, vol. XX, pp. 179–207, 1895.

An Iterative Eigensolver for Rank-Constrained Semidefinite Programming

Rajat Sanyal Aditya V. Singh Kunal N. Chaudhury

Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

Email: sanyalrajat91@gmail.com {adityavs, kunal}@iisc.ac.in

Abstract—Rank-constrained semidefinite programming (SDP) arises naturally in various applications such as max-cut, angular (phase) synchronization, and rigid registration. Based on the alternating direction method of multipliers, we develop an iterative solver for this nonconvex form of SDP, where the dominant cost per iteration is the partial eigendecomposition of a symmetric matrix. We prove that if the iterates converge, then they do so to a KKT point of the SDP. In the context of rigid registration, we perform several numerical experiments to study the convergence behavior of the solver and its registration accuracy. As an application, we use the solver for wireless sensor network localization from range measurements. The resulting algorithm is shown to be competitive with existing optimization methods for sensor localization in terms of speed and accuracy.

Index Terms—semidefinite programming, ADMM, eigensolver, convergence, registration, sensor network localization.

I. INTRODUCTION

We consider a class of rank-constrained semidefinite programs that arise in applications such as finding the largest cut in a graph [1], determining angles from their differences [2], and the registration of point clouds using rigid transforms [3]. We will focus on the registration problem, where we have N points in \mathbb{R}^d , which are divided into M overlapping point clouds. The local coordinates (and the label) of points in each point cloud are known. The task is to compute the global coordinates of all the N points [3]. We will refer to this as *rigid registration*, which has found applications in sensor network localization [4]. The maximum likelihood estimator for this problem involves optimization over rigid transforms (translations, rotations and reflections) [3]. In particular, the maximum likelihood estimate for this problem is given by the solution of

$$\min_{\mathbf{O}_1, \dots, \mathbf{O}_M \in \mathbb{O}(d)} \sum_{i,j=1}^M \text{Tr} ([\mathbf{C}]_{ij} \mathbf{O}_j^\top \mathbf{O}_i). \quad (1)$$

where $\mathbb{O}(d)$ is the set of $d \times d$ orthogonal matrices, and $[\mathbf{C}]_{ij} \in \mathbb{R}^{d \times d}$ denotes the (i, j) -th block of a certain $\mathbf{C} \in \mathbb{R}^{Md \times Md}$. We refer the reader to [3] for details. Interestingly, (1) can be seen as a non-commutative analogue of the Boolean optimization

$$\min_{x_1, \dots, x_M \in \{-1, 1\}} \sum_{i,j=1}^M c_{ij} x_j x_i,$$

which comes up in max-cut [1]. From the point of view of continuous optimization, the main challenge with (1) is that $\mathbb{O}(d)$ is not a connected manifold (apart from being nonconvex).

In the context of local optimization, this means that we cannot hope to compute the maximum likelihood estimate unless we initialize the iterations on the correct component of the domain. Since the domain in (1) has 2^M components, getting the right initialization is difficult. It was observed in [3] that if we work with the Gram matrix of $\mathbf{O}_1, \dots, \mathbf{O}_M$, then we can express (1) as a standard semidefinite program (SDP), albeit with an additional rank-constraint. The authors proposed to relax the rank constraint to obtain a standard SDP, whose solution can be computed to arbitrary precision using an interior point solver. Later, an efficient and scalable SDP solver was proposed in [4]. Though these methods can find the global minimum of the relaxed problem, there is no guarantee that the rank of the solution is exactly d (we would have solved (1) in this case); if the rank is greater than d , then the solution becomes infeasible for (1). This necessitates “rounding” of the solution of the convex relation to obtain a feasible solution for (1), which will generally be suboptimal. Our idea is to build an efficient solver that can directly tackle (1). In this regard, our contributions are as follows:

- Based on the alternating direction method of multipliers (ADMM) [5], we develop an iterative solver for (1) that involves simple updates. In particular, one of the updates is trivial, while the other is a partial eigendecomposition of a symmetric matrix.
- We show that any fixed point of our solver is a KKT point of (1). This result is novel because, as will be made explicit, a crucial assumption behind the existing analyses on nonconvex ADMM [6] [7], [8] does not hold in our case.
- We present numerical results for rigid registration and sensor network localization, which demonstrate the effectiveness of the solver in terms of accuracy and timing.

We note that ADMM based algorithms have become popular for structured convex programming [5]. Lately, the ADMM framework has been successfully applied to various nonconvex problems, even though rigorous convergence guarantees are not available. In fact, while the analysis for convex ADMM is well established, nonconvex ADMM is still a developing area of research. Some results have been reported in [6]–[8], but the analysis in these works relies on regularity assumptions on the objective that are not met for our ADMM formulation. More precisely, both updates of our ADMM solver involves constrained optimization, while at least one update in [6]–[8] is assumed to be a smooth unconstrained optimization.

The rest of the paper is as follows. In Section II, we formulate the optimization problem, based on which we develop the solver in Section III. The fixed point analysis is undertaken in Section IV. Numerical results are reported in Section V, and we conclude with a summary of the results in Section VI.

II. PROBLEM FORMULATION

It was observed in [3] that we can express (1) as a rank-constrained semidefinite program (SDP). More specifically, consider the Gram matrix \mathbf{G} of size $m \times m$, whose (i, j) -th block is $[\mathbf{G}]_{ij} = \mathbf{O}_i^\top \mathbf{O}_j$, where $i, j \in \llbracket 1, M \rrbracket$. Here and henceforth, $m = Md$, and we use $\llbracket p, q \rrbracket$ to denote the integers $\{p, \dots, q\}$. In terms of \mathbf{G} , we can reformulate (1) as

$$\begin{aligned} & \min_{\mathbf{G} \in \mathbb{S}_+^m} \quad \text{Tr}(\mathbf{C}\mathbf{G}) \\ \text{subject to} \quad & [\mathbf{G}]_{ii} = \mathbf{I}_d, \quad i \in \llbracket 1, M \rrbracket, \quad \text{rank}(\mathbf{G}) = d, \end{aligned} \quad (2)$$

where \mathbb{S}_+^m is the set of symmetric positive semidefinite matrices of size $m \times m$. This is a standard SDP, except for the additional rank constraint, which in fact makes the problem nonconvex. Following this observation, a convex relaxation (GRET-SDP) was proposed in [3] simply by dropping the rank constraint. However, the relaxation is not guaranteed to return a rank- d solution, and one is required to “round” the solution if the rank is greater than d . This can produce suboptimal solutions, i.e., the objective value of the rounded solution can be much larger than the optimum of (2).

As against this, we propose to directly tackle the original problem (1). In particular, we will demonstrate that an iterative solver can be developed for (1), where the subproblems admit closed-form solutions that can be computed efficiently. In particular, consider the variable

$$\mathbf{W} = \frac{1}{\sqrt{M}} [\mathbf{O}_1 \dots \mathbf{O}_M]^\top \in \mathbb{R}^{m \times d}.$$

Notice that we can write (1) as

$$\begin{aligned} & \min_{\mathbf{W} \in \mathbb{R}^{m \times d}} \quad \text{Tr}(\mathbf{C}\mathbf{WW}^\top) \\ \text{subject to} \quad & [\mathbf{WW}^\top]_{ii} = M^{-1}\mathbf{I}_d, \quad i \in \llbracket 1, M \rrbracket. \end{aligned} \quad (3)$$

As mentioned above, the authors in [3] choose to work with the Gram matrix \mathbf{WW}^\top . We will however continue to work with \mathbf{W} . Moreover, for reasons that will be apparent in Section III, we propose to add the redundant constraint $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_d$. That is, we replace (3) by the equivalent problem

$$\begin{aligned} & \min_{\mathbf{W} \in \mathbb{R}^{m \times d}} \quad \text{Tr}(\mathbf{C}\mathbf{WW}^\top) \\ \text{subject to} \quad & \mathbf{W}^\top \mathbf{W} = \mathbf{I}_d, \\ & [\mathbf{WW}^\top]_{ii} = M^{-1}\mathbf{I}_d, \quad i \in \llbracket 1, M \rrbracket. \end{aligned} \quad (4)$$

By “equivalent”, we mean that any optimal solution of (3) is also optimum for (4), and vice versa. Importantly, notice that unlike (2), there are no (explicit) rank constraints in (4).

III. PROPOSED SOLVER

We propose to solve (4) using variable splitting and the alternating direction method of multipliers [5]. More specifically, by introducing the variable $\mathbf{X} = \mathbf{WW}^\top$, we first transform (4) into the following problem:

$$\begin{aligned} & \min_{\mathbf{W} \in \mathbb{R}^{m \times d}, \mathbf{X} \in \mathbb{S}^m} \quad \text{Tr}(\mathbf{C}\mathbf{WW}^\top) \\ \text{subject to} \quad & [\mathbf{X}]_{ii} = M^{-1}\mathbf{I}_d, \quad i \in \llbracket 1, M \rrbracket, \\ & \mathbf{W}^\top \mathbf{W} = \mathbf{I}_d, \quad \mathbf{X} = \mathbf{WW}^\top, \end{aligned} \quad (5)$$

where \mathbb{S}^m is the set of symmetric matrices of size m . In terms of the sets

$$\Theta = \{\mathbf{W} \in \mathbb{R}^{m \times d} : \mathbf{W}^\top \mathbf{W} = \mathbf{I}_d\},$$

and

$$\Omega = \{\mathbf{X} \in \mathbb{S}^m : [\mathbf{X}]_{ii} = M^{-1}\mathbf{I}_d, i \in \llbracket 1, M \rrbracket\},$$

we can compactly write (5) as follows:

$$\begin{aligned} & \min_{\mathbf{W} \in \Theta, \mathbf{X} \in \Omega} \quad \text{Tr}(\mathbf{C}\mathbf{WW}^\top) \\ \text{subject to} \quad & \mathbf{X} = \mathbf{WW}^\top. \end{aligned} \quad (6)$$

This is a constrained optimization problem with variables \mathbf{X} and \mathbf{W} . The augmented Lagrangian for (6) is given by

$$\begin{aligned} \mathcal{L}_\rho(\mathbf{W}, \mathbf{X}, \boldsymbol{\Lambda}) = & \langle \mathbf{C}, \mathbf{WW}^\top \rangle + \langle \boldsymbol{\Lambda}, \mathbf{X} - \mathbf{WW}^\top \rangle \\ & + \frac{\rho}{2} \|\mathbf{X} - \mathbf{WW}^\top\|^2, \end{aligned} \quad (7)$$

where \mathbf{W} and \mathbf{X} are the primal variables, and $\boldsymbol{\Lambda} \in \mathbb{S}^m$ is the dual variable associated with the constraint $\mathbf{X} = \mathbf{WW}^\top$; $\rho > 0$ is a penalty parameter [5]. Notice that we have used the inner-product $\langle \mathbf{X}, \mathbf{Y} \rangle = \text{Tr}(\mathbf{XY})$ and the Frobenius norm $\|\mathbf{X}\| = \langle \mathbf{X}, \mathbf{X} \rangle^{1/2}$, both defined on \mathbb{S}^m .

Starting with some initialization $\mathbf{X}, \boldsymbol{\Lambda} \in \mathbb{S}^m$ and $\rho_0 > 0$, the ADMM iterates for $k = 0, 1, \dots$ are given by

$$\mathbf{W}^{k+1} = \arg \min_{\mathbf{W} \in \Theta} \mathcal{L}_{\rho_k}(\mathbf{W}, \mathbf{X}^k, \boldsymbol{\Lambda}^k), \quad (8)$$

$$\mathbf{X}^{k+1} = \arg \min_{\mathbf{X} \in \Omega} \mathcal{L}_{\rho_k}(\mathbf{W}^{k+1}, \mathbf{X}, \boldsymbol{\Lambda}^k), \quad (9)$$

$$\boldsymbol{\Lambda}^{k+1} = \boldsymbol{\Lambda}^k + \rho_k (\mathbf{X}^{k+1} - \mathbf{W}^{k+1} \mathbf{W}^{k+1\top}),$$

$$\rho_{k+1} = \min(\gamma \rho_k, \rho_\infty),$$

where $\gamma > 1$. Notice that ρ_k is allowed to increase at each iteration up till ρ_∞ , and then it is held fixed.

It is not difficult to verify that by combining the linear and quadratic terms, we can write (9) as

$$\mathbf{X}^{k+1} = \arg \min_{\mathbf{X} \in \Omega} \|\mathbf{X} - \mathbf{A}^k\|^2 = \mathcal{P}_\Omega(\mathbf{A}^k), \quad (10)$$

where $\mathbf{A}^k = \mathbf{W}^{k+1} \mathbf{W}^{k+1\top} - \rho_k^{-1} \boldsymbol{\Lambda}^k$, and \mathcal{P}_Ω is the projection operator. From the definition of Ω , it follows that \mathbf{X}^{k+1} is obtained simply by replacing the diagonal blocks of \mathbf{A}^k with $M^{-1}\mathbf{I}_d$, keeping the other blocks unchanged.

We next compute the update in (8). After some manipulations, we can write this as

$$\mathbf{W}^{k+1} = \arg \min_{\mathbf{W} \in \Theta} \langle \mathbf{B}^k, \mathbf{WW}^\top \rangle, \quad (11)$$

where $\mathbf{B}^k = \mathbf{C} - \boldsymbol{\Lambda}^k - \rho_k \mathbf{X}^k$. It is now clear that (11) is an eigenvalue problem. Indeed, if $\mathbf{w}_1, \dots, \mathbf{w}_d$ are the columns of \mathbf{W} , then we can write (11) as

$$\min_{\mathbf{w}_1, \dots, \mathbf{w}_d} \sum_{i=1}^d \mathbf{w}_i^\top \mathbf{B}^k \mathbf{w}_i. \quad (12)$$

Moreover, since $\mathbf{W} \in \Theta$, it follows that $\mathbf{w}_1, \dots, \mathbf{w}_d$ form an orthonormal system. Let $\lambda_1, \dots, \lambda_m$ be the eigenvalues of \mathbf{B}^k sorted in non-decreasing order, with corresponding eigenvectors $\mathbf{q}_1, \dots, \mathbf{q}_m$. Then the solution of (12) is given by $\mathbf{w}_i^* = \mathbf{q}_i, i \in \llbracket 1, d \rrbracket$. In other words,

$$\mathbf{W}^{k+1} = [\mathbf{q}_1 \cdots \mathbf{q}_d]. \quad (13)$$

The ADMM updates are summarized in Algorithm 1. The dominating cost per iteration is the (partial) eigendecomposition of \mathbf{B}^k , and this can be performed efficiently using off-the-shelf eigensolvers. Since the problem is nonconvex, the initialization of \mathbf{X} plays an important role. In this regard, we note that the eigendecomposition of \mathbf{C} can be used to obtain a non-trivial initialization of \mathbf{X} [3]. As for $\boldsymbol{\Lambda}$, we simply initialize it to zero for all the experiments.

Algorithm 1: ADMM Solver

Input: $\mathbf{C}, \gamma > 1, \rho_\infty > 0$.

Initialize: $\mathbf{X}, \boldsymbol{\Lambda}$, and $\rho > 0$.

while some stopping criteria is not met

$\mathbf{B} = \mathbf{C} - \boldsymbol{\Lambda} - \rho \mathbf{X}$.
 $\{\mathbf{q}_1, \dots, \mathbf{q}_d\}$: bottom d eigenvectors of \mathbf{B} .
 $\mathbf{W} = [\mathbf{q}_1 \dots \mathbf{q}_d]$.
 $\mathbf{X} \leftarrow \Pi_\Omega(\mathbf{W} \mathbf{W}^\top - \rho^{-1} \boldsymbol{\Lambda})$.
 $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + \rho(\mathbf{X} - \mathbf{W} \mathbf{W}^\top)$.
 $\rho \leftarrow \min(\gamma \rho, \rho_\infty)$.

end

IV. FIXED POINT ANALYSIS

Convergence analysis of ADMM for convex problems is a well-researched topic [5]. However, a theoretical understanding of why ADMM solvers applied to nonconvex programs succeed so often in practice remains elusive. Lately, there have been a handful convergence results for nonconvex ADMM [6] [7], [8]. Unfortunately, they rely on assumptions that do not hold for our problem. More precisely, note that we can rewrite (6) as

$$\begin{aligned} \min_{\mathbf{W} \in \mathbb{R}^{m \times d}, \mathbf{X} \in \mathbb{S}^m} \quad & \text{Tr}(\mathbf{C} \mathbf{W} \mathbf{W}^\top) + \iota_\Theta(\mathbf{W}) + \iota_\Omega(\mathbf{X}) \\ \text{subject to} \quad & \mathbf{X} = \mathbf{W} \mathbf{W}^\top, \end{aligned} \quad (14)$$

where ι_Γ is the indicator function associated with a feasible set Γ , namely, $\iota_\Gamma(\mathbf{Y}) = 0$ if $\mathbf{Y} \in \Gamma$, and $\iota_\Gamma(\mathbf{Y}) = \infty$ otherwise [5]. Note that, because of the indicator functions, the objective function in (14) is non-smooth in both \mathbf{W} and \mathbf{X} . This violates a crucial regularity assumption common in existing analyses of nonconvex ADMM, namely, that the objective must be smooth in at least one variable. As a result, none of the existing results on convergence are applicable to the proposed ADMM solver.

Nevertheless, we will show in Section V that Algorithm 1 performs well empirically and, in particular, it is found to converge with the spectral initialization. On the theoretical front, we have succeeded in establishing that if the iterates of Algorithm 1 converge, then they do so to a KKT (stationary) point (e.g., see [9, Chapter 3]).

Before formally stating our result, we write (1) as a nonlinear program:

$$\begin{aligned} \min_{\mathbf{O}_1, \dots, \mathbf{O}_M \in \mathbb{R}^{d \times d}} \quad & \sum_{i,j=1}^M \text{Tr}([\mathbf{C}]_{ij} \mathbf{O}_j^\top \mathbf{O}_i) \\ \text{subject to} \quad & \mathbf{I}_d - \mathbf{O}_i^\top \mathbf{O}_i = 0, \quad i \in \llbracket 1, M \rrbracket. \end{aligned} \quad (15)$$

This allows us to write the Lagrangian of (15) and use KKT theory. In particular, the Lagrangian is given by

$$\mathcal{L} = \sum_{i,j=1}^M \text{Tr}([\mathbf{C}]_{ij} \mathbf{O}_j^\top \mathbf{O}_i) + \sum_{i=1}^M \text{Tr}(\boldsymbol{\Lambda}_i (\mathbf{I}_d - \mathbf{O}_i^\top \mathbf{O}_i)), \quad (16)$$

where the symmetric matrices $\boldsymbol{\Lambda}_i \in \mathbb{R}^{d \times d}, i \in \llbracket 1, M \rrbracket$, are the Lagrange multipliers for the equality constraints (these should not be confused with the multiplier in (7)). We have the following characterization of the KKT point of (16) (see Appendix VII-A for the proof). Recall that \mathbf{G} is the Gram matrix of the \mathbf{O}_i 's, whose (i, j) -th block is $[\mathbf{G}]_{ij} = \mathbf{O}_i^\top \mathbf{O}_j$.

Lemma 1. *The variables $\mathbf{O}_1^*, \dots, \mathbf{O}_M^* \in \mathbb{R}^{d \times d}$ are a KKT point of (15) if and only if, for $i \in \llbracket 1, M \rrbracket$,*

- (a) $[\mathbf{G}^*]_{ii} = \mathbf{I}_d$, and
- (b) $[\mathbf{C}\mathbf{G}^*]_{ii} = [\mathbf{G}^*\mathbf{C}]_{ii}$.

In this case, we will say that \mathbf{G}^ is a KKT point of (15).*

We now make precise the notion of convergence that is used in our analysis. We say that Algorithm 1 has *converged* if it “stops making any progress”, i.e., the variables stop getting updated. Stated differently, if we view the progress from one iteration to the next as a map (from some space into itself), then this is equivalent to the iterates converging to a *fixed point* of this map. Indeed, note that if

$$\mathbf{W}^{k+1} \mathbf{W}^{k+1\top} = \mathbf{X}^k \quad (17)$$

for some $k = k_0$, then we must have for $k \geq k_0$:

$$\mathbf{X}^{k+1} = \mathbf{W}^{k+1} \mathbf{W}^{k+1\top} \quad \text{and} \quad \boldsymbol{\Lambda}^{k+1} = \boldsymbol{\Lambda}^k. \quad (18)$$

Conversely, if (18) holds at iteration $k = k_0$, (17) must hold for $k \geq k_0$. In summary, the convergence of Algorithm 1 is equivalent to the condition that (17) holds for some $k = k_0$. We are now in a position to state our main result (the proof is provided in Appendix VII-B).

Theorem 2. *Suppose $\boldsymbol{\Lambda}^0 = \mathbf{0}$ and $\mathbf{X}^k = \mathbf{W}^{k+1} \mathbf{W}^{k+1\top}$ at iteration $k = k_0$. Then $\mathbf{G}^* = M \mathbf{X}^{k_0}$ is a KKT point of (15).*

The practical significance of this result is that the feasibility gap $\|\mathbf{X}^k - \mathbf{W}^{k+1} \mathbf{W}^{k+1\top}\|$ can be used to monitor the evolution of the iterates to a fixed point. For example, we can stop the iterations when this gap is less than a specified tolerance.

V. NUMERICAL EXPERIMENTS

We now report some numerical experiments on rigid registration to analyze performance of the proposed solver. We also compare its performance with GRET-SDP [3], which solves a convex relaxation of (2). Moreover, as a concrete application, we use our algorithm for sensor network localization (SNL) using a registration-based framework [10]. In this context, we also compare our method with SNLSDP [11] and ESDP [12], which are still considered state-of-the-art SNL solvers as far as localization accuracy is concerned. SNLSDP solves the semidefinite relaxation of the SNL using a interior point solver [11]. Due to the poor scalability of the interior method, a further edge-based relaxation is proposed in [12] known as ESDP. The simulations were executed on a 3.4 GHz, quad-core machine with 32 GB memory, using the MATLAB implementation of Algorithm 1. For the experiments, we increase the value of ρ for the first 100 iterations, and fix it for subsequent iterations.

A. Performance analysis for rigid registration

Experiment 1. For two point clouds, i.e., when $M = 2$, the registration problem has a closed-form solution [13]. In this case, global optimum of (1) can be computed exactly. As a result, we can check optimality of the solution computed by the proposed solver for two point clouds. Specifically, we consider a point cloud with $N = 500$ points. We apply a random rigid transformation on each point cloud and corrupt the local coordinates. If the global coordinates are $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d$, then the generated local coordinates are as follows:

$$\mathbf{x}_{k,i} = \mathbf{O}_i \mathbf{x}_k + \mathbf{t}_i + \boldsymbol{\epsilon}_{k,i}, \quad \boldsymbol{\epsilon}_{k,i} \sim \mathcal{N}(\mathbf{0}, \eta \mathbf{I}_d),$$

where $\mathbf{x}_{k,i}$ is the coordinate of the k -th point in the i -th point cloud ($i = 1, 2$). A typical simulation result is shown in Fig 1. In this case, $\rho_0 = 1e-4$ and $\gamma = 1.1$, and we ran the algorithm for 500 iterations. We used a random initialization for our solver. Interestingly, the iterates converged in just two iterations, for the noiseless and noisy scenarios. The reconstructed point cloud (after alignment) and the original point cloud are shown in Fig 1. To measure the reconstruction accuracy, we use the average normalized error (ANE) [14]:

$$\text{ANE} = \left\{ \frac{\sum_{i=1}^N \|\hat{\mathbf{x}}_i - \bar{\mathbf{x}}_i\|^2}{\sum_{i=1}^N \|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_c\|^2} \right\}^{1/2},$$

where $(\hat{\mathbf{x}}_i)$ are the coordinates of the reconstructed point cloud (after alignment), $(\bar{\mathbf{x}}_i)$ are the original coordinates (ground truth), and $\bar{\mathbf{x}}_c = (\bar{\mathbf{x}}_1 + \dots + \bar{\mathbf{x}}_N)/N$ is the centroid. We notice that the proposed method can solve the registration problem exactly for any random initialization when $\eta = 0$.

Experiment 2. We now study how the proposed algorithm behaves for different values of ρ_0 . We consider the setup in Experiment 1, but we use $M = 10$ point clouds. The evolution of the objective function with iterations is shown in Fig 2, for noise level $\eta = 0.01$. Notice that the objective converges within few iterations, though it converges to different values depending on ρ_0 . In particular, the ANEs are identical ($= 0.7$) for $\rho_0 = 1e-4, 1e-3$, and $1e-2$; however, $\text{ANE} = 17.3$ when

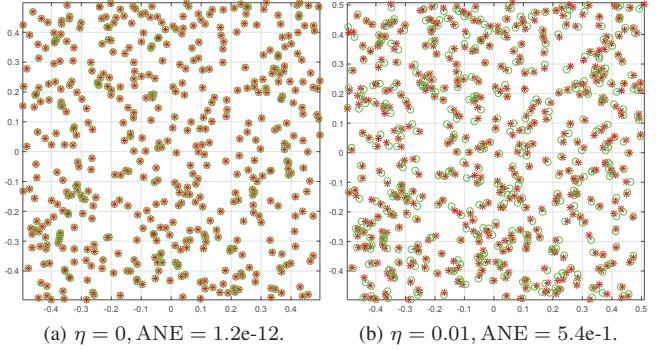


Fig. 1. Registration of two point clouds with $N = 500$ point each, both with and without noise in the local coordinates. The estimated and the original coordinates are marked using \star and \circ . We have used random initializations.

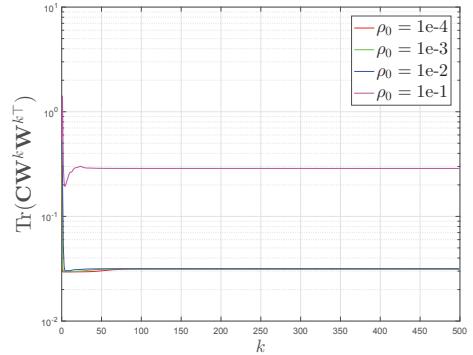


Fig. 2. Evolution of the objective function with iterations for different ρ_0 .

$\rho_0 = 0.1$ (see the plots in Fig 2). A possible explanation is that the iterates converge to a poor local minimum in the latter case. Based on exhaustive simulations, it appears that the ANE is small when ρ_0 is in the range $[1e-4, 1e-2]$. Unfortunately, unlike when $M = 2$, since we cannot ascertain the global minimum of (1) in this case, we cannot assert that the iterates converge to the optimal solution when $\rho_0 \in [1e-4, 1e-2]$.

Experiment 3. We now perform an experiment using the setup in Experiment 2, but at zero noise level. We measure the feasibility gap $\|\mathbf{X}^k - \mathbf{W}^k \mathbf{W}^{k\top}\|$ at each iteration. These are shown in Fig 3. Notice that the gap seems to vanish (up to machine precision) after a finite number of iterations. For completeness, the ANE is $9.3e-11$ (exact reconstruction).

Experiment 4. We next compare with GRET-SDP [3] in terms of timing and accuracy. We set $N = 500, M = 100, d = 2$ and $\eta = 1e-2$. For both methods, we initialize with the spectral solution GRET-SPEC [3]. Evolution of the objective is shown in Fig 4. Note that both methods converge to the same objective. However, the proposed algorithm converges much faster than the convex GRET-SDP solver. Moreover, as far as the per-iteration cost is concerned, our method requires just the bottom d eigenvectors (d is 2 or 3 for most practical problems), whereas GRET-SDP requires the full eigendecomposition. A comparison with GRET-SPEC in terms of accuracy is provided in Fig 5. Notice that the proposed method performs better than GRET-SPEC. Moreover, as with GRET-SPEC, the ANE for

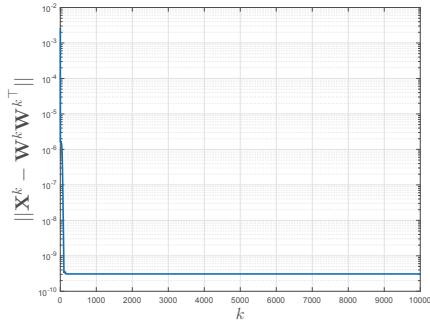


Fig. 3. Evolution of the feasibility gap with iterations.

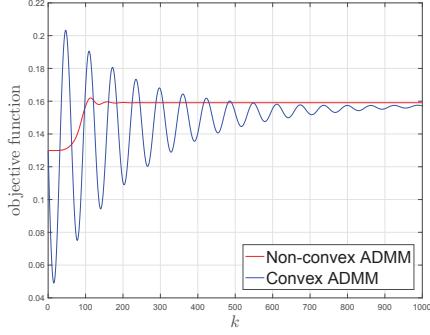


Fig. 4. Convergence results for $N = 500, M = 100, d = 2$ and $\eta = 0.01$.

our method show a linear trend with the noise level.

B. Application to sensor network localization

We now demonstrate the effectiveness of our algorithm for range-based wireless SNL. Recall that the problem in SNL is to determine the location of a network of sensors from inter-sensor distances and locations of a few selected sensors (called anchors) [11], [12]. More specifically, the distance between two sensors is assumed to be known if they are within the radio range (denoted by r) of each other. It was shown in [10] that the localization problem can be solved efficiently by mapping it into a registration problem. More specifically, the idea was to divide the wireless network into smaller overlapping cliques (wherein all the pairwise distances are known). Each clique is then efficiently localized (in parallel) using classical multidimensional scaling. Finally, the cliques are registered in a global coordinate system using rigid registration. We propose to use Algorithm (1) in place of GRET-SDP which was originally used in [10].

As for the network topology, we consider random geometric graphs (RGGs) and structured datasets. To simulate a RGG, we randomly sample points from the unit square $[-0.5, 0.5]^2$, and consider them as the sensors. We connect two sensors by an edge only if their distance is less than r . For each edge, the distance measurement is modeled as follows [11], [12]:

$$d_{ij} = |1 + \epsilon_{ij}| \cdot \|\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j\|, \quad \epsilon_{ij} \sim \mathcal{N}(0, \eta),$$

when i and j are sensors, and as

$$d_{ik} = |1 + \epsilon_{ik}| \cdot \|\bar{\mathbf{x}}_i - \mathbf{a}_k\|, \quad \epsilon_{ik} \sim \mathcal{N}(0, \eta),$$

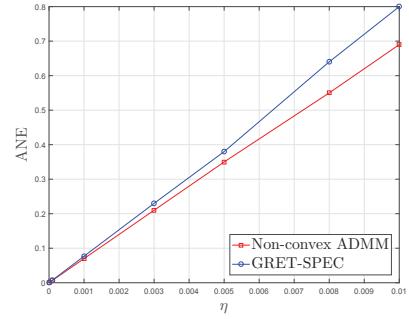


Fig. 5. Comparisons of ANEs for various η ($N = 500, M = 100, d = 2$).

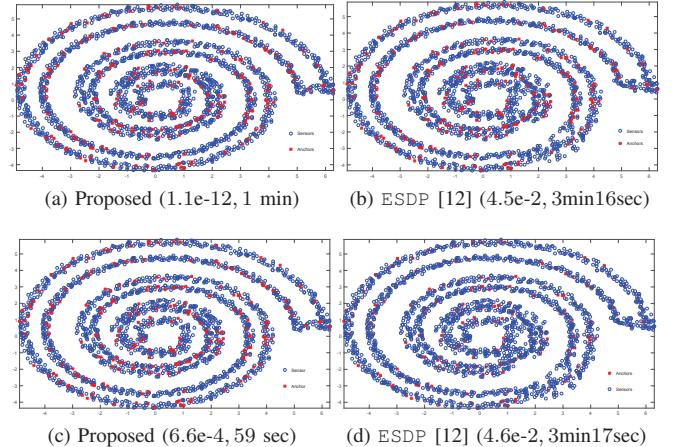


Fig. 6. Localization results for the spiral dataset [14]. The first row corresponds to clean measurements ($\eta = 0$), while the bottom row is for $\eta = 0.01$. In either case, $r = 0.8$. The anchors are shown in \star , while the localized sensors are shown in \circ . The (ANE, timing) are also reported.

if i is a sensor and k is an anchor.

Experiment 5. A detailed comparison with SNLSDP [11] and ESDP [12] is provided in Table I. For each noise level, we have averaged the results over 100 realizations. For a given N , we randomly picked 10% points and set them as anchors [11]. The ANEs and timings are compared for various network sizes and noise levels in Table I. The proposed solver can find the sensor locations at machine level precision for noiseless scenarios. Moreover, our method outperforms ESDP both in terms of the ANE and timing. Note that although the accuracy of SNLSDP and our method are comparable, SNLSDP cannot be scaled to large networks.

Experiment 6. Finally, we compare the proposed algorithm with ESDP on the spiral dataset [14]. This consists of 2259 points and its diameter (distance between two furthest points) is 11.2. We randomly select 226 points (about 10% points) as anchors, and set the radio range to $r = 0.8$. The reconstructions are reported in Fig 6, along with the corresponding ANEs and timings. Notice that, unlike ESDP, our method is able to preserve the network structure.

VI. CONCLUSION

We proposed an iterative solver for rank-constrained SDP with block diagonal constraints. The per-iteration complexity is

TABLE I
COMPARISON OF TIMING AND LOCALIZATION ACCURACY OF THE PROPOSED METHOD WITH SNLSDP [11] AND ESDP [12] FOR RANDOM GEOMETRIC GRAPHS. WE USE A \star TO MEAN THAT THE INTERIOR-POINT SOLVER COULD NOT SOLVE THE SDP IN SNLSDP FOR THAT SETTING.

N	K	r	η	Timing			Accuracy (ANE)		
				Proposed	ESDP [12]	SNLSDP [11]	Proposed	ESDP [12]	SNLSDP [11]
100	10	0.4	0	0.6sec	4.9sec	3.4sec	1.5e-14	1.8e-5	7.6e-10
			0.1	0.6sec	2.4sec	5.1sec	2.4e-2	3.2e-1	2.4e-2
500	50	0.18	0	5.6sec	1.1min	5.9min	2.5e-14	6.6e-7	6.6e-7
			0.1	5.7sec	18.7sec	8.3min	1e-2	2.3e-2	1e-2
1000	100	0.12	0	23.8sec	2.6min	*	3.5e-11	1.4e-6	*
			0.01	23sec	1.2min	*	7.7e-4	1.3e-3	*
4000	400	0.06	0	14.9min	43.2min	*	1e-13	1.8e-6	*
			0.01	14.8 min	17.8min	*	3.9e-4	1e-3	*

essentially the computation of the bottom d eigenvectors of a symmetric matrix. We proved that if the iterates converge, then they do so to a KKT point. Results of numerical simulations were reported to show that the algorithm indeed converges (and often quite rapidly) for the registration problem. Moreover, our solver was shown to compare favorably with existing methods, both in terms of the timing and accuracy. We also showed how the proposed solver can be used for sensor localization by integrating it with the registration-based framework proposed in [10]. This was shown to yield promising results (competitive with existing optimization methods [11], [12]) for both random and structured networks.

VII. APPENDIX

A. Proof of Lemma 1

For a minimization problem with equality constraints, KKT conditions amount to primal feasibility and stationarity of Lagrangian with respect to the primal variables [9]. Primal feasibility gives us condition (a). On the other hand, setting the derivative of (16) with respect to \mathbf{O}_i to zero, we obtain

$$\mathbf{O}_i \Lambda_i = \sum_{j=1}^M \mathbf{O}_j [\mathbf{C}]_{ji}, \quad i \in \llbracket 1, M \rrbracket. \quad (19)$$

Left multiplying (19) by \mathbf{O}_i^\top , we have

$$\Lambda_i = \sum_{j=1}^M \mathbf{O}_i^\top \mathbf{O}_j [\mathbf{C}]_{ji} = \sum_{j=1}^M [\mathbf{G}]_{ij} [\mathbf{C}]_{ji} = [\mathbf{GC}]_{ii}.$$

Also, note that $\Lambda_i^\top = [\mathbf{CG}]_{ii}$. Since Λ_i is symmetric, condition (b) follows immediately. Conversely, it is not difficult to see that conditions (a) and (b) together imply that \mathbf{G} is a stationary point of (15).

B. Proof of Theorem 2

To show that \mathbf{G}^* is a KKT point of (15), we use Lemma 1. Since $\mathbf{G}^* = M\mathbf{X}^{k_0}$ and $\mathbf{X}^{k_0} \in \Omega$, it is clear that $[\mathbf{G}^*]_{ii} = \mathbf{I}_d, i \in \llbracket 1, M \rrbracket$. This verifies condition (a) in Lemma 1. We now verify condition (b): $[\mathbf{CG}^*]_{ii} = [\mathbf{G}^*\mathbf{C}]_{ii}, i \in \llbracket 1, M \rrbracket$.

Since $\mathbf{G}^* = M\mathbf{X}^{k_0}$ and $\mathbf{X}^{k_0} = \mathbf{W}^{k_0+1}\mathbf{W}^{k_0+1\top}$, it follows from (13) that the eigenvectors of \mathbf{G}^* and \mathbf{B}^{k_0} are identical. Therefore, \mathbf{G}^* and \mathbf{B}^{k_0} must commute: $\mathbf{B}^{k_0}\mathbf{G}^* = \mathbf{G}^*\mathbf{B}^{k_0}$. Moreover, since $\mathbf{B}^{k_0} = \mathbf{C} - \Lambda^{k_0} - \rho_k \mathbf{X}^{k_0}$, we obtain that

$$(\mathbf{C} - \Lambda^{k_0})\mathbf{G}^* = \mathbf{G}^*(\mathbf{C} - \Lambda^{k_0}).$$

In particular, $[(\mathbf{C} - \Lambda^{k_0})\mathbf{G}^*]_{ii} = [\mathbf{G}^*(\mathbf{C} - \Lambda^{k_0})]_{ii}$. That is,

$$[\mathbf{CG}^*]_{ii} - [\Lambda^{k_0}\mathbf{G}^*]_{ii} = [\mathbf{G}^*\mathbf{C}]_{ii} - [\mathbf{G}^*\Lambda^{k_0}]_{ii}.$$

Since the update in (10) affects only the diagonal blocks and $\Lambda^0 = \mathbf{0}$ by assumption, it is easily verified that Λ^k is block diagonal for $k \geq 0$. Therefore, $[\Lambda^{k_0}\mathbf{G}^*]_{ii} = [\Lambda^{k_0}]_{ii}[\mathbf{G}^*]_{ii}$ and $[\mathbf{G}^*\Lambda^{k_0}]_{ii} = [\mathbf{G}^*]_{ii}[\Lambda^{k_0}]_{ii}$. Now, since $[\mathbf{G}^*]_{ii} = \mathbf{I}_d$, condition (b) follows.

REFERENCES

- [1] M. X. Goemans and D. Williamson, “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming,” *Journal of the ACM*, vol. 42, pp. 1115–1145, 1995.
- [2] A. Singer, “Angular synchronization by eigenvectors and semidefinite programming,” *Applied and computational harmonic analysis*, vol. 30, no. 1, pp. 20–36, 2011.
- [3] K. N. Chaudhury, Y. Khoo, and A. Singer, “Global registration of multiple point clouds using semidefinite programming,” *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 468–501, 2015.
- [4] R. Sanyal, S. M. Ahmed, M. Jaiswal, and K. N. Chaudhury, “A scalable ADMM algorithm for rigid registration,” *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1453–1457, 2017.
- [5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [6] C. Lu, J. Feng, Z. Lin, and S. Yan, “Nonconvex sparse spectral clustering by alternating direction method of multipliers and its convergence analysis,” *Proc. AAAI Conference on Artificial Intelligence*, 2018.
- [7] Y. Wang, W. Yin, and J. Zeng, “Global convergence of ADMM in nonconvex nonsmooth optimization,” *arXiv preprint arXiv:1511.06324*, 2015.
- [8] M. Hong, Z.-Q. Luo, and M. Razaviyayn, “Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems,” *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.
- [9] D. P. Bertsekas, *Nonlinear Programming*. Athena scientific Belmont, 1999.
- [10] R. Sanyal, M. Jaiswal, and K. N. Chaudhury, “On a registration-based approach to sensor network localization,” *IEEE Trans. Signal Processing*, vol. 65, no. 20, pp. 5357–5367, 2017.
- [11] P. Biswas, T.-C. Liang, K.-C. Toh, Y. Ye, and T.-C. Wang, “Semidefinite programming approaches for sensor network localization with noisy distance measurements,” *IEEE Trans. Automation Science and Engineering*, vol. 3, no. 4, pp. 360–371, 2006.
- [12] Z. Wang, S. Zheng, Y. Ye, and S. Boyd, “Further relaxations of the semidefinite programming approach to sensor network localization,” *SIAM Journal on Optimization*, vol. 19, no. 2, pp. 655–673, 2008.
- [13] K. S. Arun, T. S. Huang, and S. D. Bolstein, “Least-squares fitting of two 3-D point sets,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, no. 5, pp. 698–700, 1987.
- [14] M. Cucuringu, Y. Lipman, and A. Singer, “Sensor network localization by eigenvector synchronization over the Euclidean group,” *ACM Trans. Sensor Networks*, vol. 8, no. 3, pp. 19–42, 2012.

A weighted optimization for Fourier Ptychographic Microscopy

Parimala Kancharla

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India - 502285
ee15m17p100001@iith.ac.in*

Sumohana S. Channappayya

*Department of Electrical Engineering
Indian Institute of Technology
Hyderabad, India - 502285
sumohana@iith.ac.in*

Abstract—Fourier ptychography can be implemented as a phase retrieval optimization algorithm that iteratively solves for high resolution spectrum from low resolution images. In prior art, all the low resolution images were considered equally in the optimization. In this paper, we propose a weighted optimization algorithm to enhance the quality of reconstruction with the same convergence speed. Our method is motivated by the observation that bright field and dark field low resolution images have significantly different pixel intensities. Therefore, we weight their estimated error differently in the optimization. Though the proposed method is both conceptually and computationally simple, it dramatically improves the quality of reconstruction. We also show that the weighted optimization algorithm converges to a lower mean squared error value compared to the conventional optimization. We validate our approach on several low resolution images from an experimental dataset.

I. INTRODUCTION

Wide field of view and resolution are trade-offs in a traditional optical microscope. Fourier Ptychographic Microscopy (FPM) provides a computational solution to enhance the space bandwidth product [1] of an existing microscope via post processing. A simple LED array is introduced into a normal microscope which provides different angles of light illumination. Low resolution images are captured by sequentially lighting up the elements of the LED array. The process of acquiring low resolution images leads to a loss of phase information. This necessitates an optimization algorithm for retrieving the phase and iteratively constructing the high resolution image. The cost function for this optimization is defined to be sum of the squared error between the low resolution image and the appropriately down sampled high resolution estimate. Given the nature of this cost function, a gradient descent based algorithm is typically used to solve the problem in an iterative manner.

We review the relevant literature in the following. In their seminal work, Zheng et al. [2] solved the phase retrieval optimization [3] using the alternating projection algorithm. Yang et al. [4] derived the gradient expressions for the squared error objective function. Horstmeyer et al. [5] formulated FPM as a convex problem and derived a global minimum

solution. Their work is based on the assumption that a basis could be found where the high resolution spectrum can be expressed sparsely. However, this method under performs with respect to time complexity and noise robustness. A nonlinear convergence factor was introduced into FPM optimization and an optimum convergence factor was found empirically in Zhang et al. [6] 's work.

A comprehensive review of gradient descent methods and convex relaxation methods available to solve FPM can be found in [7]. Hao et al. [7] also validated the methods on simulation and experimental data sets of FPM. Tian et al. [8] provide a hardware solution to improve the convergence speed of FPM reconstruction. It introduced coded illumination to LED array of FPM setup. This greatly improved the speed of reconstruction as it requires capturing of few low resolution images compared to older methods. In this paper, we propose a weighted optimization for FPM, which significantly enhances the quality of reconstruction. We review the conventional FPM optimization formulation next and discuss the proposed method subsequently.

II. FOURIER PTYCHOGRAPHIC MICROSCOPY AS OPTIMIZATION

The principle of the FPM algorithm [2] is based on the assumption that angular illumination leads to a shift of the sample's spectrum, which brings a practical solution to enhance the resolution. It utilizes all the low resolution measurements to span the spectrum. The optics of the system involved gives us the relationship [2] between the spectral shift and angle of illumination, which is essential in formulating the optimization for FPM. The goal of this optimization setup is to solve for high resolution spectrum by fusing all the low resolution images in spectral domain.

A. Forward Formulation of FPM

We now review the model for the images captured through FPM setup from [2]. Consider a sample with transmission function $\Psi(r)$, where $r = (x, y)$ represents the 2D spatial coordinates in the sample plane. Suppose the sample is illuminated at a particular angle from a LED with a plane wave. The plane wave is defined by $\exp(i2\pi u_l \cdot r)$, where

$u_l = (u_{lx}, u_{ly})$ is the spatial frequency corresponding to l^{th} LED ($1 \leq l \leq N_{img}$) and N_{img} is the number of captured low resolution images. The exit wave from the sample is the product of the sample and illumination complex fields in spatial domain. The Fourier transform of this exit wave is $\Psi(u - u_l)$, which is just a shifted version of the Fourier transform of the object $\tilde{\Psi}(r)$. Where $\Psi(u) = F\{\hat{\Psi}(r)\}$ and F is the 2D Fourier transform operator.

This exit wave then low pass filtered by the pupil function $P(u)$ specific to the lens, which is usually a circle with its size defined by numerical aperture. Finally, with F^{-1} being the 2D inverse Fourier transform, we can write the intensity at the image plane as

$$I_{est}(r) = \|F^{-1}\{P(u)\Psi(u - u_l)\}\|^2. \quad (1)$$

In (1) only the magnitude part is considered because phase information is lost in the acquisition process of the FPM setup.

B. Optimization problem formulation for FPM

Most algorithms solve the FPM problem by minimizing the difference between the measured and estimated amplitude. Therefore, it can be formulated as the following optimization

$$f_A(\Psi(u)) = \sum_{l=1}^{N_{img}} \sum_{r} |\sqrt{I_l(r)} - |F^{-1}\{P(u)\Psi(u - u_l)\}\||^2, \quad (2)$$

$$\Psi(u)^* = \operatorname{argmin}_{\Psi(u)} f_A(\Psi(u)). \quad (3)$$

where $I_l(r)$ is the image acquired from the set of low resolution experimental measurement. Since the cost function $f_A(\Psi(u))$ aims to minimize the difference between the estimated amplitude and the measured amplitude over all the angles ($1 \leq l \leq N_{img}$), this is the amplitude based cost function. The cost function is solved by vectorizing the variables into 1D vectors $I_l(r)$, $\Psi(u)$. After vectorizing, $\operatorname{Diag}(P)Q_l$ will be equivalent to the pupil function $P(u)$. Q_l is the sampling matrix.

$$|g_l| = |F^{-1}\operatorname{Diag}(P)Q_l\Psi(u)| \quad (4)$$

$$\min_{\Psi(u)} f_A(\Psi(u)) = \min_{\Psi(u)} \sum_{l=1}^{N_{img}} \sum_{r} |\sqrt{I_l(r)} - |g_l||^2. \quad (5)$$

The gradient expression for (5) is given by

$$\nabla_{\Psi} f_{A,l}(\Psi) = -Q_l^T \operatorname{Diag}(\bar{P}) [F \operatorname{Diag}(\frac{\sqrt{I_l}}{|g_l|}) g_l - \operatorname{Diag}(P) Q_l \Psi]. \quad (6)$$

The update equation for the high resolution spectrum (Ψ)

$$\Psi^{(i+1,l)} = \Psi^{(i,l)} - \frac{1}{|P|^2} \nabla_{\Psi} f_{A,l+1}(\Psi^{(i,l)}), (1 \leq l \leq N_{img}), \quad (7)$$

where i indicates the iteration number. In each iteration, this update equation runs through all low resolution measurements. We obtain a local convergence point as a solution at this problem. Ψ is converted to spatial domain to get back high resolution image.

III. WEIGHTED OPTIMIZATION FORMULATION FOR FPM

We now present the proposed weighted optimization algorithm. Low resolution images can be classified as dark field images and bright field images based on numerical aperture. The motivation for our algorithm is shown in Fig. 1. It clearly illustrates the varying intensity levels of the bright field and light field low resolution images. The higher intensity in bright field images motivate us to emphasize them more in the optimization. We introduce a weighting function into the traditional optimization in (2) to normalize the intensities. The modified weighted FPM optimization objective function is

$$\min_{\Psi(u)} f_A(\Psi(u)) = \min_{\Psi(u)} \sum_{l=1}^{N_{img}} w(l) \sum_{r} |\sqrt{I_l(r)} - |g_l||^2, \quad (8)$$

where $w(l)$ is the normalized weighting function that satisfies

$$\sum_{l=1}^{N_{img}} w(l) = 1; w(l) \geq 0. \quad (9)$$

A. Weighting function

All the low resolution images are reordered based on angle of illumination. Weights for low resolution images are chosen to be proportional to the illumination numerical aperture of each corresponding LED. The only constraint on the weights is that they must satisfy (9).

In this work, we demonstrate the efficacy of the weighted approach by considering two weighting strategies: Two level weighting and Gaussian weighting. We chose these since they represent the two ends of the possible weighting strategies. In two-level weighting, the weights w_1 and w_2 that correspond to the bright field and dark field respectively are found by solving (10), and (11).

$$\sum_{l=1}^{N_{bright}} w(l) = 0.9 \quad (10)$$

$$\sum_{l=1}^{N_{dark}} w(l) = 0.1. \quad (11)$$

N_{bright} and N_{dark} are number of bright field images and dark field images respectively ($N_{image} = N_{bright} + N_{dark}$).

In the Gaussian case, the weights are found by sampling the standard normal function at N_{img} points such that (9) is satisfied. These weighting functions are illustrated in Fig.2.

B. Solution for Weighted Optimization

Weight based objective function (8) is also solved using Gerchberg Saxton Algorithm like conventional FPM. Gradient expression for the weighted objective function can easily derived as follows

$$P_{A,l}(\Psi) = -Q_l^T \operatorname{Diag}(\bar{P}) [F \operatorname{Diag}(\frac{\sqrt{I_l}}{|g_l|}) g_l - \operatorname{Diag}(P) Q_l \Psi] \quad (12)$$

$$\nabla_{\Psi} f_{A,l}(\Psi) = w(l) P_{A,l}(\Psi) \quad (13)$$

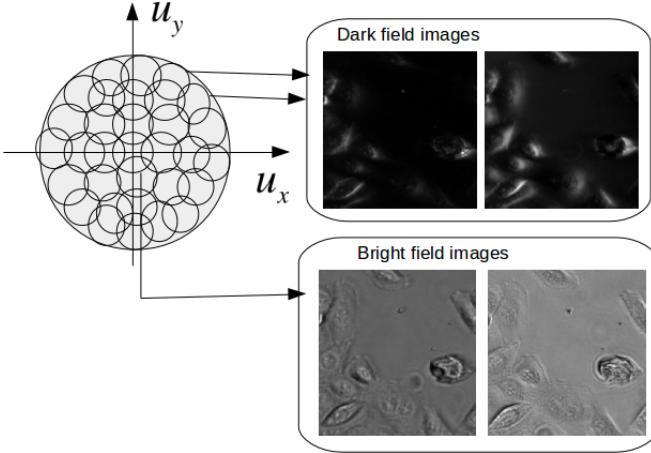


Fig.1: The sample's Fourier space is synthetically enlarged by capturing multiple images from different illumination angles. Each circle represents the spatial frequency coverage of the image captured by single LED illumination. Bright field images have more intensity compared to dark field images

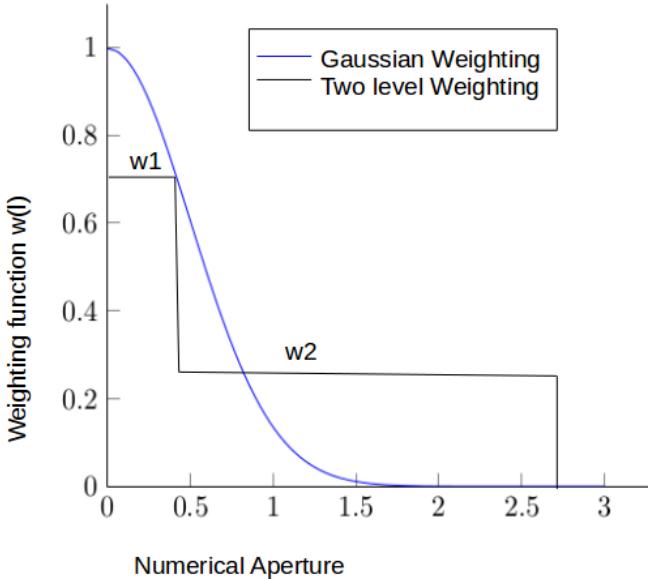


Fig.2: Weighting function

The same update equation (7) is used to reconstruct the high resolution spectrum (Ψ) with this gradient expression (12). The high resolution image is obtained by applying inverse Fourier transform to Ψ .

IV. SIMULATION RESULTS

The proposed method is validated on FPM data sets available at <http://www.laurawaller.com/opensource/>. Parameters of the experimental set up for this database are as follows. All samples are imaged with a $4 \times 0.1NA$ objective and a scientific CMOS camera. A programmable 32×32 LED array is placed

7.5mm below the sample to replace the light source on a Nikon TE300 inverted microscope. The central 293 red (central wavelength 629nm and 20nm bandwidth) LEDs are used for the experiment resulting in a final synthetic NA of 0.6.

We compared the reconstruction of conventional FPM and weighted FPM for both two level weighting and Gaussian weighting. Results show that cell structures are better resolved in case of weighted FPM compared to conventional FPM. The improvement in reconstruction quality is clearly evident in the Hela cell example Fig. 3. Another set of reconstruction results are shown for the dog tissue in Fig. 4. We have quantitatively compared the proposed method with conventional FPM by mean squared error plots. Though the proposed method is taking a few more iterations to converge, it is consistently converging to a mean squared error that is lower than the conventional method. The convergence speed of this proposed method is comparable with conventional FPM.

V. CONCLUSION

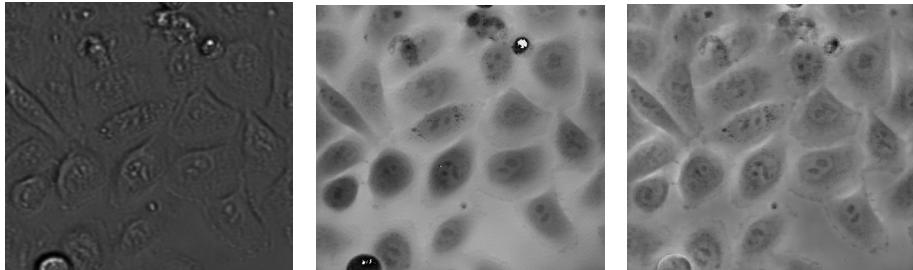
We presented a weighted optimization for solving FPM. We have evaluated its performance on open source FPM data sets. The proposed weighted optimization is simple in terms of complexity but it enhances reconstruction quality dramatically. In this work, the weights are fixed for low resolution images over iterations. In future we would like to introduce an adaptive weighting function for better error performance, where we update the weights iteratively while solving the optimization. We also would like to take weights to the pixel level making them proportional to the SNR of low resolution images to achieve noise robustness.

VI. ACKNOWLEDGEMENT

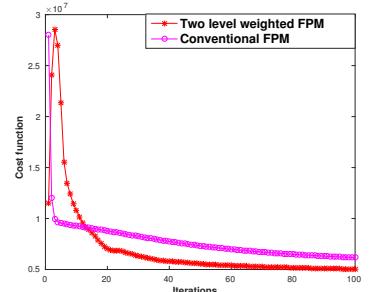
We gratefully acknowledge the support of LVPEI Hyderabad. We would like to thank Dr. Ashutosh Richchariya for his guidance.

REFERENCES

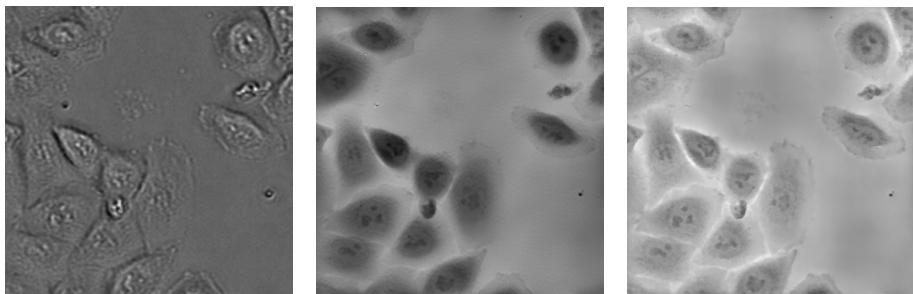
- [1] A. W. Lohmann, R. G. Dorsch, D. Mendlovic, Z. Zalevsky, and C. Ferreira, "Space-bandwidth product of optical signals and systems," *JOSA A*, vol. 13, no. 3, pp. 470–473, 1996.
- [2] G. Zheng, R. Horstmeyer, and C. Yang, "Wide-field, high-resolution fourier ptychographic microscopy," *Nature photonics*, vol. 7, no. 9, p. 739, 2013.
- [3] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.
- [4] C. Yang, J. Qian, A. Schirotzek, F. Maia, and S. Marchesini, "Iterative algorithms for ptychographic phase retrieval," *arXiv preprint arXiv:1105.5628*, 2011.
- [5] R. Horstmeyer, R. Y. Chen, X. Ou, B. Ames, J. A. Tropp, and C. Yang, "Solving ptychography with a convex relaxation," *New journal of physics*, vol. 17, no. 5, p. 053044, 2015.
- [6] Y. Zhang, W. Jiang, and Q. Dai, "Nonlinear optimization approach for fourier ptychographic microscopy," *Optics Express*, vol. 23, no. 26, pp. 33 822–33 835, 2015.
- [7] L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller, "Experimental robustness of fourier ptychography phase retrieval algorithms," *Optics express*, vol. 23, no. 26, pp. 33 214–33 240, 2015.
- [8] L. Tian, X. Li, K. Ramchandran, and L. Waller, "Multiplexed coded illumination for fourier ptychography with an led array microscope," *Biomedical optics express*, vol. 5, no. 7, pp. 2376–2389, 2014.



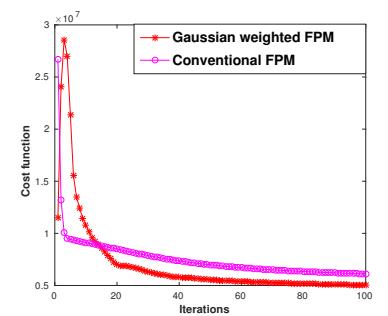
(a) Up sampled low resolution image (b) Conventional FPM reconstruction (c) Two level weighted FPM reconstruction



(d) Mean squared error plot

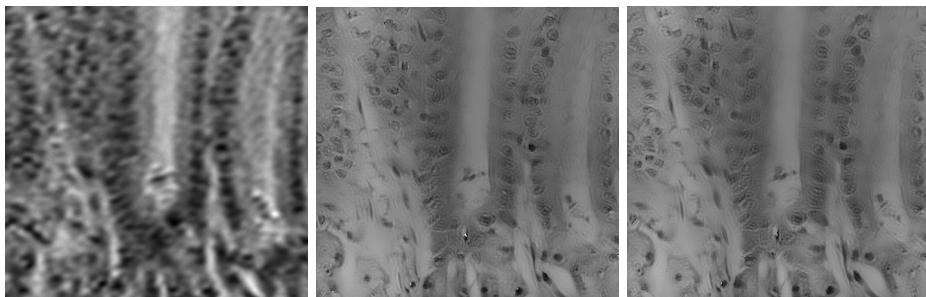


(a) Up sampled low resolution image (b) Conventional FPM reconstruction (c) Gaussian Weighted FPM reconstruction

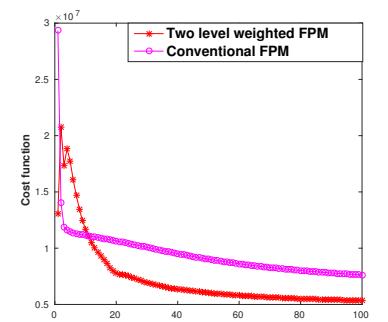


(d) Mean squared error plot

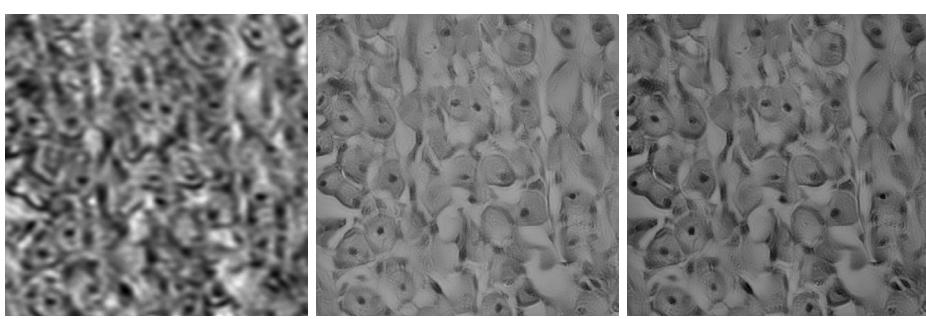
Fig.3: Hela cell reconstruction using FPM.



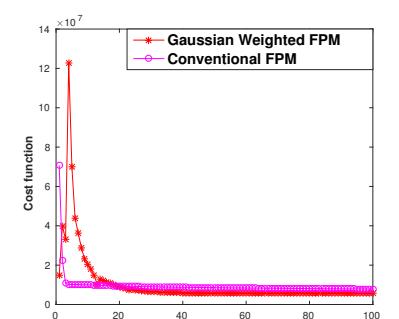
(a) Up sampled low resolution image (b) Conventional FPM reconstruction (c) Two level weighted FPM reconstruction



(d) Mean squared error plot



(a) Up sampled low resolution image (b) Conventional FPM reconstruction (c) Gaussian Weighted FPM reconstruction



(d) Mean squared error plot

Fig.3: Hela cell reconstruction using FPM.

Fig.4: Dog tissue reconstruction using FPM.

Top- m Clustering with a Noisy Oracle

Tuhinangshu Choudhury*, Dhruti Shah*, and Nikhil Karamchandani

Department of Electrical Engineering
Indian Institute of Technology, Bombay

Abstract—In this paper, we analyse the problem of top- m clustering with access to a noisy oracle. We consider a model where there are n nodes, belonging to k clusters. We have access to an oracle which when queried with a pair of nodes, returns a binary answer indicating whether they belong to the same cluster or not, but with a probability of error p . Our goal is to identify the top- m clusters in terms of size, using the noisy answers from the oracle. This setting was recently studied in [9], which provides an iterative algorithm for the case of complete clustering, i.e., $m = k$. We identify conditions (on the relative sizes of clusters) under which the first m stages of the algorithm would recover the top m clusters. We also analyze the query complexity of the algorithm and provide an upper bound which is a function of the number of recovered clusters m and the sizes of the top clusters.

I. INTRODUCTION

Clustering is one of the most basic methods of data classification and has a wide and varied history, with applications in recommendation engines, web search, social network analysis etc. This paper deals with a variant of the classical clustering problem, where we have access to a noisy oracle. A formal model for such a setting was recently introduced in [9], which considers a set of n nodes which belong to k apriori unknown clusters. There exists an oracle which accepts queries of the form “Do nodes u and v belong to the same cluster?”. The oracle provides a binary answer to the query, which is assumed to be incorrect with some probability $p < 1/2$. Thus, the answer from the oracle is strictly better than a random guess. Furthermore, it is assumed that the answer remains the same if the same pair (u, v) is queried multiple times. However, we can sequentially query the oracle with distinct pairs of nodes and then use the results to design a clustering algorithm. Such a setting can be used to model various scenarios such as crowdsourced entity resolution [11] and signed edge prediction [12], and is intimately connected to correlation clustering [8] and the Stochastic Block Model [1]. Similar noisy oracle models have also been used in other applications, for example sorting from noisy queries [3].

The goal in [9] is to characterize the minimum number of oracle queries needed to correctly identify all the k clusters, with high probability as the number of nodes n grows large. The paper provides an algorithm for this task and derives an upper bound on the total number of queries

needed, which is shown to be $O\left(\frac{nk \log n}{(1-2p)^2}\right)$. Furthermore, they provide an information-theoretic lower bound of $\Omega\left(\frac{nk}{(1-2p)^2}\right)$ on the query complexity, thus demonstrating the approximate optimality of their proposed algorithm.

This work studies a variant of the above problem where the goal is to only identify the top m clusters in terms of their size. There are several scenarios where the clusters are asymmetric in terms of size and it might suffice for an application as long as we can identify the largest few clusters. An example of such a scenario would be an online forum which answers questions relating to software bugs. In this case, the questions posted by users form the nodes. Two nodes will be considered to belong to the same cluster if they are related to the same software bug. A lot of these questions can be related to the same popular software bugs, and we would like to merge these together, while the less popular software bugs would have very few questions. Here we might be interested in finding the cluster of questions corresponding to the more popular software bugs, i.e., the top few largest clusters.

The algorithm proposed in [9] for complete clustering works in stages, and identifies one new cluster in each stage. We consider the same algorithm for the case of top- m clustering and identify conditions (on the relative sizes of clusters) under which the first m stages of the algorithm would recover the top m clusters. We also analyze the query complexity of the algorithm and provide an upper bound which is a function of the number of recovered clusters m and the sizes of the top clusters.

II. THE MODEL

Consider a set V containing n nodes, bundled in k clusters V_1, V_2, \dots, V_k . There is an oracle with error parameter $p < 1/2$. The oracle takes as input a pair of nodes and returns $+1$ if the queried nodes belong to the same cluster and -1 if they belong to different clusters, with probability of error p in both cases. Assuming $|V_1| \geq |V_2| \geq |V_3| \geq \dots \geq |V_k|$, we are interested in finding the m largest clusters, V_1, V_2, \dots, V_m .

Our algorithm assumes that V_1, V_2, \dots, V_m are of size atleast $O\left(\frac{\log n}{(1-2p)^2}\right)$. Since we are interested in finding only the larger clusters this assumption is valid. We also assume that repeated queries for same pair of nodes will give the same answer. This is to avoid cases when you repeatedly query a pair to identify whether they belong to same cluster or not.

The main result of our paper is the following theorem, which gives an upper bound on the query complexity. In the following sections we state our algorithm and prove its correctness and establish a bound on the number of pair-wise queries required, proving the theorem.

*These two authors have equal contribution.

This work was supported in part by the Bharti Centre for Communication at IIT Bombay. The work of Nikhil Karamchandani was supported in part by an Indo-French grant on “Machine Learning for Network Analytics” and the INSPIRE Faculty Fellowship from the Govt. of India.

Theorem 1. There exists an algorithm with query complexity $\min\{A, B\}$, where

$$A = O\left(\sum_{i=1}^m \min\left\{\frac{n}{|V_i|}c \log n \left(1 + \frac{1}{\log(\log n)}\right), k c \log n (1 + \delta)\right\}^2 + nm c \log n\right)$$

$$B = O\left(\frac{nk \log n}{(1 - 2p)^2}\right),$$

and $c = \frac{16}{(1-2p)^2}$, $\delta > 0$, which returns the top- m clusters with high probability, when $\frac{|V_{m+1}|}{|V_m|} \leq \left(1 - \sqrt{\frac{3 \log(k \log n)}{c \log n}}\right)$ holds.

Note that the term B in the statement above corresponds to the order-optimal query complexity of recovering all the k clusters [9]. To compare our result with the one given in [9], consider the following example scenario. Say the largest cluster is of size $O(n/3)$ while all the other clusters are of size $o(\sqrt{n})$. Thus, the total number of clusters $k = \Omega(\sqrt{n})$. Let p be some constant and say we are interested in extracting only the largest cluster, i.e., $m = 1$. In this case, we see that the terms A and B are approximately $O(n \log n)$ and $O(nk \log n)$ respectively. Thus, A is order-wise much smaller than B . From our result above, the query complexity of recovering the largest cluster is at most A while that for recovering all the k clusters is order-wise at least B [9].

III. THE ALGORITHM

We borrow our algorithm from [9] and modify it to extract only the m -largest clusters, rather than all the clusters. Our algorithm comprises of two phases. In Phase 1, we build a graph where nodes are picked randomly from V and edge weights assigned by the oracle. We then extract a subset of the largest cluster. In Phase 2, we grow this extracted subset, to include all the remaining nodes belonging to that cluster. This is repeated m times, to get the top m clusters.

Algorithm: Initialize G' with a randomly selected node from V .

1) Phase 1 Selecting a small subgraph

- Pick an unassigned node randomly from V and query it with all nodes already in G' , where edge weights (+1 or -1) between two nodes are decided by the oracle's answer.
- Extract the subgraph from G' having the maximum total edge weight, i.e. the maximum weighted subgraph (MWS) from G' and call this S .
- If $|S| \geq c \log n$, for $c = \frac{16}{(1-2p)^2}$, move onto Phase 2 else move to step (a).

2) Phase 2 Growing the extracted subcluster

- Select an unassigned vertex v not in G' , and for the MWS S extracted in the previous phase, pick $c \log n$ distinct nodes and query v with them. If the majority of these answers are positive, include v in S . Repeat this step for all unexplored nodes not in G' .

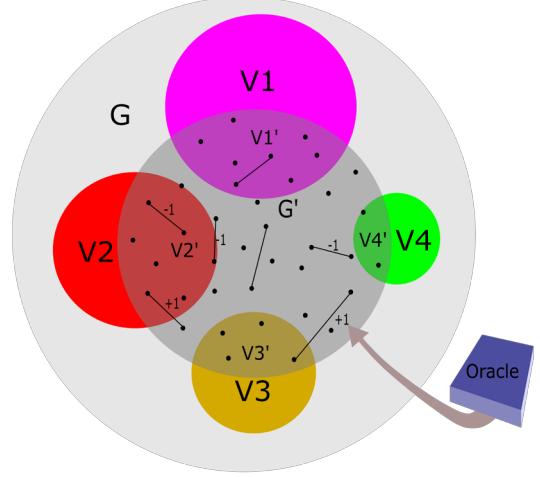


Figure 1. Graphical representation of Phase 1. V'_i denotes the subset of nodes of V_i present in the graph G' . Edges of G' are formed according to the answers obtained from the oracle.

- When there is no new vertex to query, include S in the list grown of fully grown clusters.
- Update G' by removing S from it, and every edge incident on S .
- Go back to Phase 1, and repeat the process till grown has m clusters.

At the end of the algorithm, the m clusters in grown are the top- m clusters, for $\frac{|V_{m+1}|}{|V_m|} \leq \left(1 - \sqrt{\frac{3 \log(k \log n)}{c \log n}}\right)$, as shown in Section IV.

IV. ANALYSIS

To establish the correctness of the algorithm, we show that in each of the m iterations of the algorithm, we get one of the m largest clusters. i.e. in Phase 1 we obtain as the MWS, a subset of one of the m largest clusters, which is fully grown and added to grown at the end of Phase 2. So, performing m iterations, we get all the top m clusters. We show the above for $m = 1$, i.e. we recover the largest cluster, and the same argument follows for any $m > 1$. To show this, we use the following line of argument.

- A subcluster V'_i in G' is a subset of V_i . Lemma 1 shows that when we move from Phase 1 to Phase 2, S will be one of the subclusters in G' . Further, Lemma 2 shows that S , which is a subcluster, is the one corresponding to the largest cluster, i.e. the S is V'_1 . Hence, the MWS S , of size at least $c \log n$ extracted in Phase 1, is the subcluster corresponding to the largest cluster V_1 . Lastly, Lemma 4 sets an upper bound on the total number of nodes queried in Phase 1, which is useful to obtain the query complexity of this phase.
- Lemma 3 shows that all nodes belonging to the largest cluster V_1 , and nodes from no other cluster are present in S at the end of Phase 2.

All of the following Lemmas hold true with probability approaching one as n grows large.

Lemma 1. For an MWS S of graph G' , where $|S| \geq c \log n$, S is one of the subclusters of G' .

Proof. The lemma is proved via a series of claims. The proofs of the claims are delegated to Appendix VII-A. The following claim 1 follows from [9] with minor modifications to obtain a tighter bound.

Claim 1. Let S be the MWS of G' . If

$|S| \geq c\sqrt{1-2p}\log n$, for $c = \frac{16}{(1-2p)^2}$, then S cannot contain nodes from multiple subclusters.

Claim 2. Consider a graph $\mathcal{G}(\hat{V}, \hat{E})$, where $|\hat{V}| \geq c\log n$ and edge weights w_{ij} are IID random variable taking the value -1 with probability $p < \frac{1}{2}$ and 1 with probability $1-p$. Then, $\text{weight}(\mathcal{G}) > \text{weight}(\text{any subgraph of } \mathcal{G})$ i.e. the MWS extracted, will be the entire node set V .

From Claim 1 we get that the MWS cannot contain nodes from multiple subclusters, and from Claim 2 we get that the MWS cannot be a subset of a subcluster, where $|S| \geq c\log n$. Hence, S is indeed one of the subclusters. \square

Lemma 1 ensures that when we extract an MWS from G' which is of size at least $c\log n$, then it is sure to be a subcluster of G' with high probability. The following Lemma 2 shows that for certain conditions on $\frac{|V_2|}{|V_1|}$, the MWS extracted at the end of Phase 1 will be the subcluster corresponding to the largest cluster, i.e. V'_1 .

Lemma 2. Consider $k = o(n)$. For

$\frac{|V_2|}{|V_1|} \leq \left(1 - \sqrt{\frac{4\log(k\log n)}{c\log n}}\right)$, at the end of Phase 1 the size of G' will be at most $\left(1 + \frac{1}{\log(\log n)}\right) \frac{n}{|V_1|} c\log n$ and V'_1 will be extracted as the MWS.

Proof. The lemma is proved via a series of claims. The proofs of the claims are delegated to Appendix VII-B.

Claim 3. If we sample $t_0 = \left(1 + \frac{1}{\log(\log n)}\right) \frac{n}{|V_1|} c\log n$ nodes, then $|V'_1| \geq c\log n$ and $|V'_i| \leq c\log n$, $\forall i \neq 1$.

Claim 3 implies that $|V'_1|$ will be of size at least $c\log n$ and consequently Claim 2 ensures that the weight of V'_1 will be greater than the weight of any of its own subsets. The following claim shows that the weight of V'_1 will be greater than the weight of any subset of any other subcluster, and hence ensures that V'_1 will be the MWS.

Claim 4. If we sample $t_0 = \left(1 + \frac{1}{\log(\log n)}\right) \frac{n}{|V_1|} c\log n$ nodes, the weight of V'_1 is greater than the weight of any other subset of subcluster in G' , i.e. the MWS is V'_1 .

Claim 4 ensures with high probability that the MWS is V'_1 and Claim 3 ensures that it is of size atleast $c\log n$, implying that Phase 1 ends. The Claims also ensure that the size of G' will be at most $t_0 = \left(1 + \frac{1}{\log(\log n)}\right) \frac{n}{|V_1|} c\log n$, by the time Phase 1 ends. For a general $m > 1$, a similar bound can be established on $\frac{|V_{m+1}|}{|V_m|}$, which ensures that an MWS corresponding to one of the top- m clusters is obtained at the end of Phase 1. \square

Thus, Lemma 1 and Lemma 2 show the correctness of Phase 1 of the algorithm. The following Lemma 3 is borrowed from [9], and establishes the correctness of Phase 2 of the algorithm and ensures that the subcluster obtained at the end of Phase 1 is grown fully. \square

Lemma 3. Majority of queries in Phase 2 with $c\log n$ nodes of V'_1 returns $+1$ for $v \in V_1$, and -1 for $v \notin V_1$.

Upto now we have shown the correctness of our algorithm, which returns the largest cluster when iterated over once. For a general $m > 1$, repeating the algorithm m times gives the top- m clusters. We now show the bounds on the query complexity of the algorithm.

V. QUERY COMPLEXITY

Lemma 4. The number of pair-wise queries required in a single iteration of Phase 1 of the algorithm are $\min\left\{\frac{n}{|V_1|} c\log n \left(1 + \frac{1}{\log(\log n)}\right), kc\log n(1+\delta)\right\}^2$.

Proof. The first part of the expression follows from Lemma 2. For the second part of the above expression, assume Phase 1 is not complete , when number of nodes t in G' is greater than $kc\log n(1+\delta)$. Now $t > kc\log n(1+\delta)$ implies there is atleast one subcluster S_1 of size greater than $c\log n(1+\delta)$. By Hoeffding inequality we get the weight of $S_1 \geq (1+\delta)^2(1-\epsilon)\frac{c^2\log^2 n}{2}(1-2p)$, for $\delta > 0, \epsilon > 0$, with high probability as n becomes large. This implies that the MWS will have size at least $c\log n\sqrt{(1-2p)(1+\delta)^2(1-\epsilon)}$. Choosing appropriate δ, ϵ we can have size of MWS at least $c\log n\sqrt{1-2p}$ which implies from Claim 1 that the MWS can contain nodes from single subcluster only. Claim 2 ensures that weight of S_1 is greater than the weight of any subset of S_1 . Now consider a subcluster S_2 of size less than $c\log n$.

$$\begin{aligned} & \mathbb{P}\left(\sum_{i,j \in S_1} w_{ij} < \sum_{i,j \in S_2} w_{ij}\right) \\ &= \mathbb{P}\left(\sum_{i,j \in S_1} w_{ij} - \sum_{i,j \in S_2} w_{ij} < 0\right) \\ &\leq \exp\left(-\frac{1}{3}\mathbb{E}\left(\sum_{i,j \in S_1} w_{ij} - \sum_{i,j \in S_2} w_{ij}\right)\right) \\ &\leq \exp\left(-\frac{1}{6}(1-2p)(|S_1|^2 - |S_2|^2)\right) \\ &\leq \exp\left(-\frac{1}{6}(2\delta + \delta^2)\log^2 n\right) \\ &= 0 \text{ as } n \rightarrow \infty \end{aligned}$$

This shows that weight of S_1 is greater than weight of S_2 . Similar calculation applies even when S_2 is a subset of subcluster. Hence, we get weight of S_1 is greater than any subset of subcluster of size less than $c\log n$. Hence MWS can contain nodes from a subcluster whose size is at least $c\log n$ and by Claim 2 the MWS will then be the whole subcluster itself. Hence, the MWS will be a subcluster of size atleast $c\log n$ implying Phase 1 is complete. Hence we have a contradiction.

Also, noting the bound on t_0 from Lemma 2, we get that G' will reach the size of at most

$$\binom{\mathcal{N}}{2} \leq \min\left\{\frac{n}{|V_1|} c\log n \left(1 + \frac{1}{\log(\log n)}\right), kc\log n(1+\delta)\right\}^2$$

\square

Lemma 4 gives the query complexity for a single iteration of Phase 1. Meanwhile, in the second phase we pair-wise query the unassigned nodes with $c \log n$ nodes of the MWS, so this phase requires at most $nc \log n$ queries in each iteration. To extract the top m clusters, these steps are repeated m times. So, we get an overall query complexity of

$$A = O\left(\sum_{i=1}^m \min \left\{ \frac{n}{|V_i|} c \log n \left(1 + \frac{1}{\log(\log n)}\right), k c \log n (1 + \delta) \right\}^2 + nm c \log n\right)$$

where $c = \frac{16}{(1-2p)^2}$. Also, result from [9] shows that query complexity to recover all the clusters is $B = O(\frac{nk \log n}{(1-2p)^2})$. Hence query complexity is $\min\{A, B\}$.

VI. CONCLUSION

In this paper we have analysed the problem of extracting the top few largest clusters from the query results of a noisy oracle. We have provided an algorithm for the same, and shown its correctness under certain assumptions. We also provide an upper bound on the the query complexity of the algorithm. While the proposed algorithm gives the correct clustering results, it is not computationally efficient when the number of nodes is large. This can be partially addressed using some of the ideas presented in [9]. Also, it would be interesting to derive information-theoretic lower bounds on the number of queries for the case of top- m clustering, along the same lines as those provided in [9] for the case of complete clustering.

REFERENCES

- [1] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS)*, pages 670–688, 2015.
- [2] Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In *Advances in neural information processing systems*, pages 3216–3224, 2016.
- [3] Mark Braverman and Elchanan Mossel. Noisy sorting without resampling. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 268–276. Society for Industrial and Applied Mathematics, 2008.
- [4] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.
- [5] William Feller, RH Fox, DC Spencer, AW Tucker, C Kuratowski, W Sierpiński, F Hausdorff, R von Mises, and L Pontryagin. Vol. 1,0 an introduction to probability theory and its applications. *Bull. Amer. Math. Soc.*, 1968.
- [6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.
- [7] David R Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in neural information processing systems*, pages 1953–1961, 2011.
- [8] Claire Mathieu and Warren Schudy. Correlation clustering with noisy input. In *Symposium on Discrete Algorithms (SODA)*, pages 712–728, 2010.
- [9] Arya Mazumdar and Barna Saha. Clustering with noisy queries. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5788–5799, 2017.
- [10] Arya Mazumdar and Barna Saha. Query complexity of clustering with side information. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4682–4693, 2017.
- [11] Arya Mazumdar and Barna Saha. A theoretical analysis of first heuristics of crowdsourced entity resolution. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 970–976, 2017.

- [12] Charalampos E Tsourakakis, Michael Mitzenmacher, Kasper Green Larsen, Jarosław Błasiok, Ben Lawson, Preetum Nakkiran, and Vasileios Nakos. Predicting positive and negative links with noisy queries: Theory & practice. *arXiv preprint arXiv:1709.07308*, 2017.

VII. APPENDIX

A. Proof of claims in Lemma 1

Proof of Claim 1. The following proof follows from [9] with minor modifications to obtain a tighter bound.

Let $S \not\subseteq V_i$ for all i . Then S must have intersection with at least 2 clusters. Let $V_i \cap S = C_i$ and let $j^* = \arg \min_{i: C_i \neq \emptyset} |C_i|$.

We claim that

$$\sum_{i,j \in S, i < j} w_{ij} < \sum_{i,j \in S \setminus C_{j^*}, i < j} w_{ij}$$

with high probability. The statement explains that weight of set is less if it contains nodes from multiple subclusters.

The above condition is equivalent to

$$\sum_{i,j \in C_{j^*}, i < j} w_{ij} + \sum_{i \in C_{j^*}, j \in S \setminus C_{j^*}} w_{ij} < 0$$

We will use the following inequalities mentioned in the appendix of [4] in the proof

$$\mathbb{P}(X \geq (1 + \epsilon)\mathbb{E}(X)) \leq \exp\left(-\frac{\epsilon^2}{3}\mathbb{E}(X)\right) \quad (\text{VII-A.1})$$

$$\mathbb{P}(X \leq (1 - \epsilon)\mathbb{E}(X)) \leq \exp\left(-\frac{\epsilon^2}{3}\mathbb{E}(X)\right) \quad (\text{VII-A.2})$$

where $X = \sum_i X_i$, X'_i 's are IID

$$1) |C_{j^*}| > c(1 - 2p)^{\frac{3}{2}} \sqrt{\log(\log n) \log n}, \quad c = \frac{16}{(1-2p)^2}.$$

Now

$$\begin{aligned} \mathbb{P}\left(\sum_{i,j \in C_{j^*}, i < j} w_{ij} > (1 + \nu)(1 - 2p)\binom{|C_{j^*}|}{2}\right) \\ \stackrel{(a)}{\leq} \exp\left(-\frac{\nu^2}{3}(1 - 2p)\frac{|C_{j^*}|^2}{4}\right) \\ \leq \exp\left(-\frac{16^2\nu^2}{12}\log(\log n) \log n\right) \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{P}\left(\sum_{\substack{i \in C_{j^*} \\ j \in S \setminus C_{j^*}}} w_{ij} > -(1 - \nu)(1 - 2p)|C_{j^*}||S \setminus C_{j^*}|\right) \\ \stackrel{(b)}{\leq} \exp\left(-\frac{\nu^2}{3}(1 - 2p)|C_{j^*}|^2\right) \rightarrow 0 \end{aligned}$$

Here (a), (b) follows from VII-A.1. Hence

$$\begin{aligned} \sum_{i,j \in C_{j^*}, i < j} w_{ij} + \sum_{i \in C_{j^*}, j \in S \setminus C_{j^*}} w_{ij} \\ \leq (1 + \nu)(1 - 2p)\frac{|C_{j^*}|^2}{2} - \\ (1 - \nu)(1 - 2p)|C_{j^*}||S \setminus C_{j^*}| \\ \stackrel{(c)}{\leq} (1 + \nu)(1 - 2p)\frac{|C_{j^*}|^2}{2} - \\ (1 - \nu)(1 - 2p)|C_{j^*}||C_{j^*}| < 0 \end{aligned}$$

with high probability as $n \rightarrow \infty$. Here (c) follows from the fact that C_{j^*} is the smallest subcluster in S , hence $|C_{j^*}| \leq \frac{|S|}{2} \Rightarrow |S \setminus C_{j^*}| \geq |C_{j^*}|$

2) $|C_{j^*}| < c(1-2p)^{\frac{3}{2}}(\sqrt{\log(\log n) \log n})$. Using above mentioned inequalities,

$$\begin{aligned} & \mathbb{P}\left(\sum_{\substack{i \in C_{j^*} \\ j \in S \setminus C_{j^*}}} w_{ij} > -(1-\nu)(1-2p)|C_{j^*}||S \setminus C_{j^*}|\right) \\ & \stackrel{(a)}{\leq} \exp\left(-\frac{\nu^2}{3}(1-2p)|C_{j^*}||S \setminus C_{j^*}|\right) \\ & \stackrel{(b)}{\leq} \exp\left(-\frac{\nu^2}{3}(1-2p)(|S|-1)\right) \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

where (a) follows from VII-A.1 and (b) follows from the fact that $-|C_{j^*}||S \setminus C_{j^*}|$ takes the maximum value at $|C_{j^*}| = |S| - 1$. Also let $|C_{j^*}| = x$. Hence

$$\sum_{i,j \in C_{j^*}, i < j} w_{ij} \leq \frac{x^2}{2}$$

Therefore the sum becomes

$$\begin{aligned} & \sum_{i,j \in C_{j^*}, i < j} w_{ij} + \sum_{i \in C_{j^*}, j \in S \setminus C_{j^*}} w_{ij} \\ & \leq \frac{x^2}{2} - (1-\nu)(1-2p)|C_{j^*}||S \setminus C_{j^*}| \\ & \leq \frac{x^2}{2} - (1-\nu)(1-2p)x(|S| - x) \\ & \leq x\left(\frac{3x}{2} - (1-\nu)(1-2p)|S|\right) \\ & \stackrel{(a)}{\leq} x\left(\frac{3c(1-2p)^{\frac{3}{2}}(\sqrt{\log(\log n) \log n})}{2} - (1-\nu)(1-2p)c(\sqrt{1-2p} \log n)\right) \\ & \leq c(1-2p)^{\frac{3}{2}}x\left(\frac{3(\sqrt{\log(\log n) \log n})}{2} - (1-\nu) \log n\right) \stackrel{(b)}{<} 0 \end{aligned}$$

where (a) follows from the fact that $|S| > c\sqrt{1-2p} \log n$; (b) is true for large enough n

Thus S will consist of nodes from a single sub-cluster only w.h.p. as $n \rightarrow \infty$ \square

Proof of Claim 2. For this part, we will use the following inequalities

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} < n! < e n^{n+\frac{1}{2}} e^{-n} \quad (\text{VII-A.3})$$

$$\mathbb{P}(X \leq \mathbb{E}(X) - t) \leq \exp\left(-\frac{t^2}{2n}\right) \quad (\text{VII-A.4})$$

where $X = \sum_{i=1}^n X_i$, and X_i 's are IID. Equation VII-A.3 is a direct consequence of stirling approximation mentioned in [5] and VII-A.4 is from appendix of [4]. Let S be a subset of \hat{V} . We will try to find the probability that weight of S is greater than \hat{V}

$$\begin{aligned} & \mathbb{P}\left(\sum_{i,j \in \hat{V}, i \neq j} w_{ij} < \sum_{i,j \in S, i \neq j, S \subseteq \hat{V}} w_{ij}\right) \\ & = \mathbb{P}\left(\sum_{(i,j) \in (\hat{V}, \hat{V}), (i,j) \notin (S, S), i \neq j} w_{ij} < 0\right) \\ & \stackrel{(a)}{\leq} \exp\left(-2(1-2p)^2\left[\binom{|\hat{V}|}{2} - \binom{|S|}{2}\right]\right) \end{aligned}$$

(a) follows from VII-A.4. Applying the union bound gives

$$\begin{aligned} & \mathbb{P}(MWS \neq \hat{V}) \\ & = \sum_{|S|=1}^{|\hat{V}|-1} \binom{|\hat{V}|}{|S|} \mathbb{P}\left(\sum_{(i,j) \in (\hat{V}, \hat{V}), (i,j) \notin (S, S), i \neq j} w_{ij} < 0\right) \\ & \leq \sum_{|S|=1}^{|\hat{V}|-1} \binom{|\hat{V}|}{|S|} \exp\left(-2(1-2p)^2\left(\binom{|\hat{V}|}{2} - \binom{|S|}{2}\right)\right) \\ & \leq \sum_{|S|=1}^{|\hat{V}|/2} \binom{|\hat{V}|}{|S|} \exp\left(-2(1-2p)^2\left(\binom{|\hat{V}|}{2} - \binom{|S|}{2}\right)\right) + \\ & \quad \sum_{\substack{|S|=1 \\ |\hat{V}|/2+1}}^{|\hat{V}|-1} \binom{|\hat{V}|}{|S|} \exp\left(-2(1-2p)^2\left(\binom{|\hat{V}|}{2} - \binom{|S|}{2}\right)\right) \\ & \stackrel{(b)}{\leq} \sum_{|S|=1}^{|\hat{V}|/2} \binom{|\hat{V}|}{|S|} \exp\left(-(1-2p)^2\left(\frac{3|\hat{V}|^2}{4} - \frac{|\hat{V}|}{2}\right)\right) + \\ & \quad \sum_{|S|=|\hat{V}|/2+1}^{|\hat{V}|-1} \binom{|\hat{V}|}{|S|-1} \exp\left(-2(1-2p)^2(|\hat{V}|-1)\right) \\ & \stackrel{(c)}{\leq} \frac{|\hat{V}|}{2} \frac{e}{\pi} \frac{2^{|\hat{V}|}}{\sqrt{|\hat{V}|}} \exp\left(-(1-2p)^2\left(\frac{3|\hat{V}|^2}{4} - \frac{|\hat{V}|}{2}\right)\right) \\ & \quad + \frac{|\hat{V}|}{2} |\hat{V}| \exp\left(-2(1-2p)^2(|\hat{V}|-1)\right) \\ & \leq k' \sqrt{|\hat{V}|} \exp\left(|\hat{V}|(\log 2 + 1) - (1-2p)^2 \frac{3|\hat{V}|^2}{4}\right) \\ & \quad + \frac{k''}{2} |\hat{V}|^2 \exp(-16 \log n) \\ & \stackrel{(d)}{\leq} k' \sqrt{n} \exp\left(\frac{16 \log 2 e}{(1-2p)^2} \log n - \frac{192}{(1-2p)^2} \log^2 n\right) \\ & \quad + \frac{k''}{2} \frac{n^2}{n^{16}} \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

(b): For the 2nd term, we will prove that the maximum value correspond to $|S| = |\hat{V}|-1$; (c): the 1st term follows from VII-A.3 (d) follows from the fact that the function is decreasing and $|\hat{V}| \geq c \log n$, hence $|\hat{V}| = c \log n$ gives the upper bound. Also k' and k'' are constants

To prove that second term maximizes when $|S| = |\hat{V}|-1$. Consider the function

$$\begin{aligned} f(a) & = \binom{|\hat{V}|}{|\hat{V}|-a} e^{-2(1-2p)^2\left(\binom{|\hat{V}|}{2} - \binom{|\hat{V}|-a}{2}\right)} \\ & = \binom{|\hat{V}|}{|\hat{V}|-a} e^{-(1-2p)^2(2a|\hat{V}|-a^2+a)} \end{aligned}$$

Clearly the $f(a)$'s denote the terms of the second expression in the above proof. We want to show that the maxima occurs at $a = 1$. Now

$$\begin{aligned} & \frac{f(a)}{f(a+1)} \\ & = \frac{\binom{|\hat{V}|}{|\hat{V}|-a} e^{-(1-2p)^2(2a|\hat{V}|-a^2+a)}}{\binom{|\hat{V}|}{|\hat{V}|-a-1} e^{-(1-2p)^2(2(a+1)|\hat{V}|-(a+1)^2+a+1)}} \end{aligned}$$

$$\begin{aligned}
&= \frac{\frac{|\hat{V}|!}{a!(|\hat{V}|-a)!} e^{-(1-2p)^2(2a|\hat{V}|-a^2+a)}}{\frac{|\hat{V}|!}{(a+1)!(|\hat{V}|-a-1)!} e^{-(1-2p)^2(2(a+1)|\hat{V}|-(a+1)^2+a+1)}} \\
&= \frac{a+1}{|\hat{V}|-a} \frac{e^{2(1-2p)^2|\hat{V}|}}{e^{(1-2p)^2(2a+2)}} \\
&\stackrel{(a)}{\geq} \frac{2}{|\hat{V}|-1} \frac{e^{2(1-2p)^2|\hat{V}|}}{e^{(1-2p)^2(2(\frac{|\hat{V}|}{2}-1)+2)}} \\
&\geq \frac{2}{n} e^{(1-2p)^2|\hat{V}|} \stackrel{(b)}{\geq} 2n^{15} \geq 1
\end{aligned}$$

(a) follows from mimimizing individual fraction when $a \in [0, \frac{|\hat{V}|}{2} - 1]$: (b) follows from the fact that $|\hat{V}| \geq \frac{16}{(1-2p)^2} \log n$. Hence $f(a)$ is a decreasing function of a , implying maxima occurs at $a = 1$. \square

B. Proof of claims in Lemma 2

Proof of Claim 3. Consider that we sample $t_0 = \left(1 + \frac{1}{\log(\log n)}\right) \frac{n}{|V_1|} c \log n$ nodes.

$$\mathbb{E}[|V'_1|] = \frac{|V_i|}{n} t_0$$

Using Chernoff bound we get the following,

$$\begin{aligned}
&\mathbb{P}(|V'_1| \leq (1-\epsilon)\mathbb{E}[|V'_1|]) \leq \exp\left(-\frac{\epsilon^2 \mathbb{E}[|V'_1|]}{2}\right) \\
\implies &\mathbb{P}\left(|V'_1| \leq (1-\epsilon)\frac{|V_1|}{n} t_0\right) \leq \exp\left(-\frac{\epsilon^2}{2} \frac{|V_1|}{n} t_0\right)
\end{aligned}$$

Choose $\epsilon = 1 - \frac{n}{|V_1|} \frac{c \log n}{t_0}$ to get

$$\begin{aligned}
&\mathbb{P}(|V'_1| \leq c \log n) \\
&\leq \exp\left(-\frac{\left(1 - \frac{n}{|V_1|} \frac{c \log n}{t_0}\right)^2 |V_1|}{2} t_0\right) \\
&\leq \exp\left(-\frac{1}{2(1+\log(\log n)) \log(\log n)} c \log n\right) \rightarrow 0
\end{aligned}$$

Hence if we sample $t_0 = \left(1 + \frac{1}{\log(\log n)}\right) \frac{n}{|V_1|} c \log n$ nodes then $|V'_1| \geq c \log n$ with high probability. Similarly,

$$\begin{aligned}
&\mathbb{E}(|V'_2|) = \frac{|V_2|}{n} t_0 \\
&= \left(1 + \frac{1}{\log(\log n)}\right) \frac{|V_2|}{|V_1|} c \log n \\
&\leq c \log n \left(1 + \frac{1}{\log(\log n)}\right) \left(1 - \sqrt{\frac{4 \log(k \log n)}{c \log n}}\right) \\
&\leq c \log n \left(1 - \sqrt{\frac{3 \log(k \log n)}{c \log n}}\right)
\end{aligned}$$

$$\mathbb{P}(|V'_2| \geq c \log n)$$

$$\begin{aligned}
&\leq \mathbb{P}\left(|V'_2| \geq \mathbb{E}(|V'_2|) \left(1 + \left(\frac{c \log n}{\mathbb{E}(|V'_2|)} - 1\right)\right)\right) \\
&\leq \exp\left(-\frac{1}{3} \left(\frac{c \log n}{\mathbb{E}(|V'_2|)} - 1\right)^2 \mathbb{E}(|V'_2|)\right)
\end{aligned}$$

$$\begin{aligned}
&\leq \exp\left(-c \log n \frac{\left(1 - \left(1 - \sqrt{\frac{3 \log(k \log n)}{c \log n}}\right)\right)^2}{3\left(1 - \sqrt{\frac{3 \log(k \log n)}{c \log n}}\right)}\right) \\
&\leq \exp\left(-c \log n \frac{\log(k \log n)}{c \log n}\right) \leq \frac{1}{k \log n}
\end{aligned}$$

Union bound gives

$$\mathbb{P}(|V'_i| \geq c \log n) \leq k \frac{1}{k \log n} = \frac{1}{\log n} \rightarrow 0 \text{ as } n \rightarrow \infty$$

Hence $|V'_i| \leq c \log n$, $\forall i \neq 1$, when we query t_0 nodes. \square

Proof of Claim 4.

$$\mathbb{E} \sum_{\substack{s,t \in S \\ s < t}} w_{s,t} = \binom{|S|}{2} (1-2p)$$

Using Hoeffding's inequality [6], we have

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\substack{s,t \in V'_1 \\ s < t}} w_{s,t} < \sum_{\substack{s,t \in S \\ s < t}} w_{s,t}\right) \\
&\leq \exp\left(-\frac{(1-2p)^2 \left[\binom{|V'_1|}{2} - \binom{|S|}{2}\right]^2}{2 \left[\binom{|V'_1|}{2} + \binom{|S|}{2}\right]}\right)
\end{aligned}$$

Taking union bound $\forall S \subseteq V'_1$ and $\forall i$, we get

$$\begin{aligned}
&\mathbb{P}\left(\sum_{\substack{s,t \in V'_1 \\ s < t}} w_{s,t} < \sum_{\substack{s,t \in S \\ s < t}} w_{s,t}\right) \\
&\leq k 2^{|V'_2|} \exp\left(-\frac{(1-2p)^2 \left[\binom{|V'_1|}{2} - \binom{|V'_2|}{2}\right]^2}{2 \left[\binom{|V'_1|}{2} + \binom{|V'_2|}{2}\right]}\right) \\
&\approx k 2^{|V'_2|} \exp\left(-(1-2p)^2 (|V'_1| - |V'_2|)^2\right) \\
&\leq k 2^{(1-\delta')|V'_2|} \exp\left(-(1-2p)^2 [(1-\delta)\mathbb{E}|V'_1| - (1+\delta')\mathbb{E}|V'_2|]\right) \\
&\leq k 2^{\left(1 + \frac{1}{\log(\log n)}\right) \frac{|V'_2|}{|V_1|} c \log n} \exp\left(-(1-2p)^2 \times \left(1 + \frac{1}{\log(\log n)}\right)^2 c^2 \log^2 n \right. \\
&\quad \left. \times \left[(1-\delta) - \frac{|V_2|}{|V_1|} - \beta \sqrt{\frac{|V_2|}{|V_1|}}\right]^2\right)
\end{aligned}$$

For this error to go to zero with increasing n , we get the following condition

$$|V_2| \leq \frac{|V_1|}{\beta^2} \left(1 - \delta - \frac{1}{\sqrt{\log n}}\right)^2$$

and

$$\beta < \sqrt{\frac{|V_1|}{|V_2|}} - \sqrt{\frac{|V_2|}{|V_1|}}$$

Choosing $\beta = 1 - \delta$, we have

$$|V_2| \leq |V_1| \left(1 - \frac{1}{(1-\delta)\sqrt{\log n}}\right)^2$$

Hence if $\frac{|V_2|}{|V_1|} \leq \left(1 - \sqrt{\frac{4 \log(k \log n)}{c \log n}}\right)$, the above condition is already satisfied. \square

Unsupervised GIST based Clustering for Object Localization

Saprem Shah, Kunal Khatri, Purva Mhasakar
*Information and Communication Technology,
Dhirubhai Ambani Institute of Information
and Communication Technology, India*
{201401107, 201501011, 201601082}@daiict.ac.in

Rajendra Nagar and Shanmuganathan Raman
*Electrical Engineering,
Indian Institute of Technology Gandhinagar,
Gujarat, India*
{rajendra.nagar,shanmuga}@iitgn.ac.in

Abstract—In the past years, there have been several attempts for the task of object localization in an image. However, most of the algorithms for object localization have been either supervised or weakly supervised. The work presented in this paper is based on the localization of a single object instance, in an image, in a fully unsupervised manner. Initially, from the input image, object proposals are generated where the proposal score for each of these proposals is calculated using a saliency map. Next, a graph by the GIST feature similarity between each pair of proposals is constructed. Density-based spatial clustering of applications with noise (DBSCAN) is used to make clusters of proposals based on GIST similarity, which eventually helps us in the final localization of the object. The setup is evaluated on two challenging benchmark datasets - PASCAL VOC 2007 dataset and object discovery dataset. The performance of the proposed approach is observed to be comparable with various state-of-the-art weakly supervised and unsupervised approaches for the problem of localization of an object.

Index Terms—Object Localization, Unsupervised Learning, GIST, DBSCAN

I. INTRODUCTION

Object localization is an important and highly challenging problem faced in the field of computer vision where the main aim is to figure out the location as well as estimating a bounding box around the different categories of objects present in an image. It serves as a crucial and one of the primary steps where machines need to understand an image deeper, segment it and finally recognize the objects present in the image. Consider Figure 1. Most people, in spite of pictures having variations as well as different backgrounds, can easily understand it and label it in one category. This is because humans can easily focus on relevant objects, declutter unnecessary background and cluster images belonging to the same class. However, this interpretation proves to be a major hurdle to computers because of the high intra-class variations, occlusion, scale variations, background clutter, and viewpoint variations. Thus object localization has become a challenging problem in computer vision.

Since many years, sliding window approach ([9], [17], [32], [33]) has been the most common method used in order to search for an object in the image. Despite being successful, this approach proves to be costly because a sequential system has to classify a huge number of windows for a specific input image. Not only do we need to search all possible locations



Fig. 1. Images with different background clutter and orientations but same target object

in the image, but search at different scales too. Being inefficient in speed, many subwindow search algorithms were introduced [1]. Although they performed quite well, these methods needed manually-annotated bounding boxes to be around the objects of all categories present in an image. This labeling of a large amount of data accurately and manually is a very expensive and cumbersome task and hence, weakly-supervised learning methods have been proposed ([18], [28], [35], [36], [38]). These use only category labels during training. That is, they require negative and positive image-level labels for an object class under target, making the data manageable. Thus, rather than focusing on the bounding box annotations of the objects, it is focused on image level labeling. However, even these techniques require human efforts. For providing a fundamental solution to annotation dependency, unsupervised object localization techniques have started emerging. Among the less supervised approaches, co-localization and co-segmentation based algorithms have emerged. Both the algorithms in spite of relaxing the need of annotations still require a set of images which contain the common dominant object class which is to be localized and thus imports some sort of supervision too. Unlike most of the present algorithms, we have presented a completely unsupervised algorithm for object localization of a single object present in an image without any target object or assumptions in the process. This makes the entire problem more challenging and important.

In this paper, we advocate an unsupervised approach to object localization by initially extracting object proposals and then filtering out candidate regions by using saliency map. We then model a graph based on GIST descriptors. Both from literature and our experience, global features like GIST have proved to be more robust, less complex and more computationally efficient than the local feature detectors for image localization, scene categorization, and

scene change detection [40]. This is supported by the fact that the identification of an object by a human eye global recognition is first cared before going into details. This perceptual information of the scene understood by the human eye is the GIST. Based on the perceptual similarity, the proposals are clustered using DBSCAN. DBSCAN is not restricted to spherical clusters. It is noise resistant and is able to handle outliers. It captures complex shapes well and thus shows better results for object localization than other clustering methods like k -Means Clustering. Mean of top-scoring proposals is taken to get the final localized window.

The main contributions of this paper are the following:

- A completely unsupervised algorithm based on GIST similarity features is introduced.
- Clustering of localized proposals which makes use of data clustering algorithm DBSCAN with appropriate parameters is proposed for estimating final localized window.
- A simple but efficient algorithm for object localization is proposed and explored on challenging benchmark datasets such as the object discovery dataset [16] and the PASCAL VOC 2007 dataset [14].

II. RELATED WORK

Object localization and discovery using unsupervised learning has been attempted a few times in computer vision. The results of most of these approaches were much less accurate than their counterpart - weakly supervised and supervised algorithms. Due to the difficulty of fully unsupervised learning techniques for localization, many past and recent works are focused on weakly supervised and supervised approaches. One of the most commonly used object localization methods has been using sliding window classifiers. There are many existing sliding window approaches ([4], [9], [20], [32]). The classifier is subsequently applied to sub-images, thus obtaining a classification map as well as the region with the maximum score. Testing every possible image location by scanning over the entire test image becomes computationally inefficient and costly. Blaschko *et al.* in [3] improvised the existing approach by proposing a two-stage approach. In this method, scoring was done using non-linear (SVM) classifier and the linear support vector machine (SVM) classifier was used for pre-selection. Due to the high cost of data-labeling, it is difficult to obtain strong supervision information, thus weakly supervised learning gained increasing attention. Weakly supervised learning can be divided into four categories: (i) multiple instance learning [37], [44], (ii) Exhaustive search technique [30], (iii) inter-intra-class modeling [12], and (iv) topic model [39].

Recently, more work has been directed towards a weakly supervised approach from different angles. Co-segmentation was first proposed by Rother *et al.* by merging the Markov random fields with color histogram matching to segment objects common to two images [31]. This approach has been improvised and extended to handle more general cases ([19], [26], [42], [43]). Co-localization has the same type of input as co-segmentation ([15], [27], [29]). However, co-segmentation

segments out common foreground regions from a given set of images while co-localization seeks localized objects with bounding boxes. Tang *et al.* presented a joint image-box formulation method for solving the co-localization problem [41]. Moreover, a co-localization method built on clustering of local features along with partial correspondences was presented by Grauman and Darrell [19]. Recently, many co-localization methods based on pre-trained deep convolutional models have been discovered. However, almost all co-localization methods cannot handle noisy data. Some of the methods extracted the fully connected representations considering the pre-trained models as feature extractors. Recently, there have been few works which are based on completely unsupervised fashion ([7], [45]). Vora *et al.* proposed an iterative spectral clustering approach [45]. They have formulated the task as an undirected graph using HOG descriptor [11] and performed iterative spectral clustering on the graph constructed. Unlike most of the previous works, our setup is completely unsupervised. We propose to perform object localization of a single object without any target object in the given input image.

III. PROPOSED APPROACH

A. Saliency

Selective search which is an off-the-shelf algorithm has been used initially for extracting initial object proposal from a given input image [42]. Prospective objects from this image are identified using segmentation. The algorithm gives bounding boxes as an output corresponding to all patches in an image which is most probable to be the objects. After the extraction of initial object proposals, $P = \{p_1, p_2, \dots, p_N\}$ from the image I , where $N = 1000$, object proposals are further processed as following. Each of the initial extracted proposal is scored by the probability that the area enclosed by it contains an object. Further saliency scores of each proposal are calculated which helps in filtering the initial proposals based on certain aspects of the image. For calculating this score, we have computed a saliency map M of the input image I which is primarily based on the saliency algorithm presented by Margolin *et al.* [27]. Thus by saliency scores calculated for each proposal, the proposals covering the prominent regions of the image, i.e., those having high proposal score, form our initial set of proposals. The saliency map M of the image I is of size $w \times h$ and is used to determine the score for each proposal. We calculate the sum of pixel values of all the pixels present in the proposal, let us call this value total sum ts . Next, we divide the ts by $\sqrt{h \times w} \times \max(h, w)$ to get the final proposal score for a proposal S_p .

$$S_p = \frac{ts}{\sqrt{h \times w} \times \max(h, w)}$$

Here, we do not simply take the average of all pixel values contained in an object proposal because even an object proposal covering just a sub-part of an object can have a high score in this scenario. Instead, the denominator is formulated in this manner : $\sqrt{h \times w} \times \max(h, w)$. As,

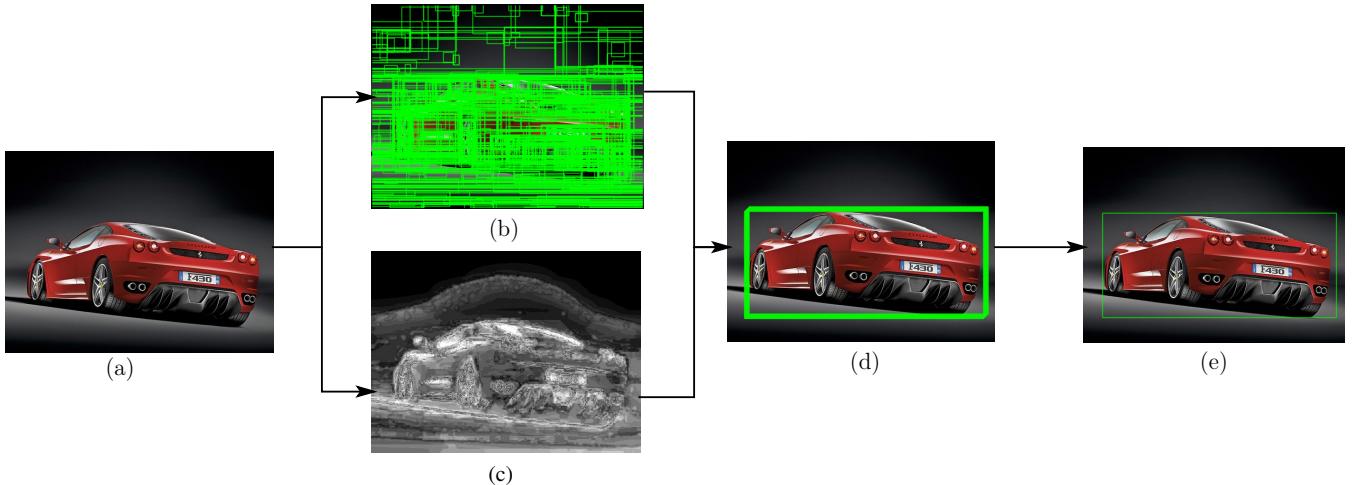


Fig. 2. Pipeline of the proposed approach. (a) Input image. (b) Extracting object proposals from the image using Selective Search algorithm (N). (c) Computing the saliency map of the image. After computing the saliency map, we compute the scores for each of the object proposals. (d) Applying Gist feature detector over all the proposals and using DBSCAN until the number of proposals becomes less than the stopping criteria (T). (e) Taking mean of all the remaining proposals to estimate the final window P_{final}

here, even the size of the proposal is taken into account. Here, those proposals will have a high score which contain a larger part of an object (probably even the entire object) than those just containing a sub-part of it.

B. GIST

After performing the process defined above, we get a set of proposals, $P = \{p_1, p_2, \dots, p_N\}$ and proposal scores, $S = \{s_1, s_2, s_3, \dots, s_N\}$, where $N = 1000$. However, it is quite possible that many of these proposals are similar, overlap or do not contain an object. Of all these proposals, we choose a subset which has the highest probability of having an object present in the image. For this, we model an undirected graph by the similarity in features among the object proposals. We extract a GIST descriptor g_i [29] for each of the proposal p_i in P . GIST provides us with a low dimensional representation of the image, without using any kind of segmentation. The GIST features, thus extracted, provide a rough representation (the GIST) of the image based on the gradient data, orientations, and scales for different areas of the image. We model the graph D by the GIST feature similarity between each pair of proposals, computed in set P . The graph contains the computed GIST descriptor g_i as the node and weighted edges. The weight of each edge of the graph is given by D_{ij} . D_{ij} stands for the Gaussian similarity score which is computed as $D_{ij} = \exp(-\frac{\|g_i - g_j\|^2}{\sigma^2})$. For the experiments conducted in this paper, σ is chosen as $0.05 \times \max(g_i - g_j)$.

C. DBSCAN

Now, we apply DBSCAN [13] to make clusters of proposals based on GIST similarity. DBSCAN groups together data points that are nearby based on a distance measurement (say Euclidean distance) and a minimum

number of points are required to make a dense region while simultaneously marking the points as outliers that are in low-density regions. It helps us in finding associations and structures in data which on the other hand are hard to find manually. These associations are useful in finding the patterns in the image. We apply several iterations of DBSCAN as a single iteration does not select highly localized object proposals. At each step, we change the input parameter, maximum distance between two points in a cluster, to DBSCAN by dividing the previous one by 2. For cluster selection as we partition subsequently, a cluster score is computed which is the average of proposal scores s of all the proposals which are there in the cluster. We discard all the lower scored proposals present in clusters. We continue this until we meet the stopping criteria; the number of proposals becomes less than or equal to 10. Finally, for obtaining the final localization window P_{final} , we consider the mean of all the coordinates of the clustered proposals obtained after DBSCAN in this final set. This is the predicted bounding box by our algorithm.

IV. EVALUATION

The setup is evaluated on two datasets, PASCAL VOC 2007 dataset [14] and Object Discovery Dataset [34]. We have compared our results with various state-of-the-art algorithms ([10], [18], [21]) which include various weakly supervised, co-localization, and co-segmentation algorithms and some completely unsupervised algorithms ([8], [45]) for object localization. The results of other methods are taken from their respective papers. We set the various parameters of our algorithm as follows: (a) Total number of proposals for each image, $T = 1000$, (b) Stopping criteria for DBSCAN, number of remaining proposals ≤ 10 , and (c) The maximum distance between 2 points in a cluster in DBSCAN is halved every iteration and initialized to 1. All

Algorithm 1: Pipeline for Object Localization

Input: Image I
Result: Final localization window P_{final}

1 Procedure:

- 2 Set the number of total proposals N .
- 3 Set the stopping condition for the algorithm, number of remaining proposals, T .
- 4 Initialize max_dist parameter for DBSCAN.
- 5 Extract the object proposals $P = \{p_1, p_2, \dots, p_N\}$ for image I .
- 6 Compute Saliency map M for image I
- 7 Compute proposal scores, $S = \{s_1, s_2, s_3, \dots, s_N\}$ and saliency scores $x = \{x_1, x_2, \dots, x_N\}$ for each generated proposal.
- 8 Compute overall score $o_i = s_i * x_i$, $i = 1$ to N .
- 9 Compute GIST features for each object proposal $G = \{g_1, g_2, \dots, g_N\}$.
- 10 **while** $N \leq T$ **do**
- 11 Find GIST feature similarity graph matrix D .
- 12 Apply DBSCAN on this graph.
- 13 Calculate cluster scores and select the cluster with the maximum score
- 14 Update proposal set P .
- 15 Reduce max_dist to half.
- 16 **end**
- 17 Compute mean of coordinates of the remaining proposals to predict the final object localization window P_{final} .

the parameters mentioned above have been found by trying out several different values and the ones above give the best results for our algorithm. These values are kept constant and used throughout the paper. We have used CorLoc (correct location) metric, which has been used by various previous works on weakly supervised and unsupervised algorithms, to measure the percentage of images correctly classified. Here, an object is considered to be correctly localized if the value of intersection-over-union score of predicted bounding box, b_p and the ground truth bounding box, b_g , is more than 0.5, that is, $\frac{\text{area}(b_p \cap b_g)}{\text{area}(b_p \cup b_g)} > 0.5$. This metric for evaluation was suggested by Everingham et al. [15]. We face a problem when there are multiple instances of objects in an image since a single object is localized in our image algorithm. This issue arises when an image contains multiple objects and since a lot of such images are present in PASCAL VOC 2007 dataset [14], it becomes challenging. We evaluate images on per image basis. That is, we consider an image to be correctly localized even if only one of the objects present in the image has been localized correctly and has satisfied the CorLoc condition. This assumption is similar to [46].

A. Computation Time

We perform our experiments on a computer, having Intel Core i7 processor. The algorithm localizes an object from an image of resolution 500 x 400 (from PASCAL VOC dataset) in about 20s. We can improve the run-time of our algorithm

by replacing our object proposal algorithm (Selective Search) with other faster object proposal algorithms ([6], [47]).

B. Object discovery dataset



Fig. 3. Predicted localized boxes on Object Discovery Dataset. Green Box = Ground Truth and Blue Box = Estimated Box

We evaluate our algorithm on Object Discovery Dataset [34], which contains images from 3 categories - car, airplane, and horse. This dataset has been widely used to benchmark algorithms for object discovery ([5], [16], [34]). Our algorithm is evaluated on 100 image subsets of each of these categories (car, airplane, and horse). As each of the images of this dataset consists of just one object, this dataset can provide a much more effective and extensive evaluation of our algorithm. Until now, none of the previously applied weakly supervised algorithms ([10], [18], [21], [28], [37], [38], [46]) have used this dataset to benchmark their evaluations. This dataset contains some noisy images without the query object, so we have to discard the noisy images for our evaluation as the ground truths for these images are not available. The airplane class has 18 outlier images, the car class has 11 outlier images and the horse class has seven outlier images. Each class contains a total of 100 images. These outlier images are noisy images as they contain no object belonging to their category and hence need to be removed before evaluation. Hence, we are left with a total of 264 (82+89+93) images from the whole dataset on which we evaluate our algorithm. The ground truths for each of these images are available in the form of segmentation, so we use these segmented images, containing the ground truth, and convert them into localization boxes. Finally, after this pre-processing, the algorithm is evaluated on all the 264 images, which are spread over all the three classes, present in the dataset. The images of our results obtained by localization on the Object Discovery Dataset are shown in Figure 3. We conduct separate-class experiments as well as mixed-class experiment on a collection of all the three classes which contains 36 outlier images. The separate class results are quite comparable and better than most co-segmentation algorithms ([22], [23], [25]). Comparison of our results of the localization of images on Object Discovery dataset is shown

in Table I. In spite of being fully unsupervised unlike most of the co-segmentation and co-localization algorithms, our algorithm gives significantly good results with the CorLoc measure of **56.29%**. Our localization results for mixed-class setup is almost similar to the separate class setup with a CorLoc measure of **54.81%**. The class-wise results for mixed-class are shown in Table II.

TABLE I
CORLOC (%) ON SEPARATE CLASS ON OBJECT DISCOVERY DATASET.

Methods	Airplane	Car	Horse	Avg
Kim et al. [25]	21.95	0.00	16.13	12.69
Joulin et al. [22]	32.93	66.29	54.84	51.35
Vora et al. [45]	43.9	65.17	45.16	51.41
Joulin et al. [23]	57.32	64.04	52.69	58.02
Tang et al. [41]	71.95	93.26	64.52	76.58
Rubinstein et al. [34]	74.39	87.64	63.66	75.16
Ours	47.56	69.66	51.61	56.29

TABLE II
CORLOC (%) ON MIXED CLASS ON OBJECT DISCOVERY DATASET.

Airplane	Car	Horse	Average
47.56	67.41	49.46	54.81

C. PASCAL VOC 2007 dataset

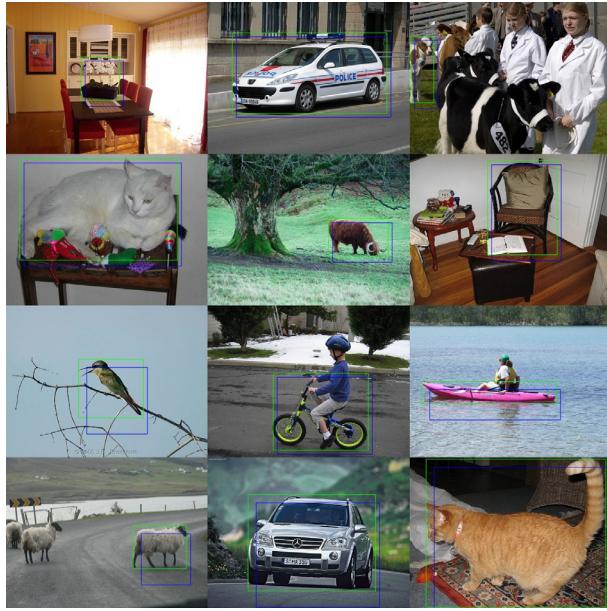


Fig. 4. Predicted localized boxes on PASCAL VOC 2007 dataset. Green Box = Ground Truth and Blue Box = Estimated Box

The PASCAL VOC 2007 dataset in [14] is much more challenging compared to the Object Discovery dataset due to significant occlusion, clutter, and contains real-life images with diverse viewpoints. Thus it is a task on a much larger scale. We compare our algorithm on the PASCAL VOC 2007 dataset with other weakly supervised and unsupervised algorithms. The dataset contains 20 object classes. We take the entire test dataset consisting of 4952 images and all the

different object classes are spread over these 4952 images. Several images contain multiple objects, ranging from dogs to humans, but we have considered our evaluation on a single object per image basis, that is, an image is correctly localized if any of the objects present in the image has been successfully localized. The results by various algorithms on the PASCAL VOC 2007 dataset are tabulated in Table III and comparisons are made with other algorithms. The data used by the mentioned algorithms are shown in the second column. The positive images in the training set are indicated by P while the negative images are indicated by N. We observe that weakly-supervised methods use more training data in comparison to unsupervised algorithms. Also, we must consider that the best performing method [46] uses additional supervised data (A) - CNN features which are pre-trained on the ImageNet dataset. Our algorithm gives CorLoc measure of **43.21%** which is better than almost all weakly supervised techniques [2], [18], [28], [35], [37], and also quite comparable to the best performing method [46] without any sort of supervision and extra manual labor of annotation or using transfer learning to extract feature descriptors. Next,

TABLE III
PERFORMANCE OF ALGORITHM ON PASCAL VOC 2007 DATASET.

Methods	Data Used	CorLoc (%)
Nguyen et al. [28]	P+N	22.4
Joulin et al. [24]	P	24.6
Andrews et al. [2]	P+N	25.4
Siva and Xiang [38]	P+N	30.2
Siva et al. [36]	P+N	30.4
Vora et al. [45]	-	35.08
Shi et al. [35]	P+N	36.2
Cho et al. [7]	-	36.6
Gokberk et al. [18]	P+N	38.6
Cinbis et al. [10]	P	47.3
Wang et al. [46]	P+N+A	48.5
Ours	-	43.21

to check the performance of the intermediate stage of our algorithm, we directly take the proposals, after computing the saliency map, having the highest proposal scores (here 10 such proposals are taken). We then take the mean of the coordinates of these proposals and output the result as the final localization window. The results are shown in Table IV.

TABLE IV
PERFORMANCE OF EACH STAGE OF OUR ALGORITHM ON PASCAL VOC 2007 DATASET.

Stages	Top-10 proposals	Overall
CorLoc (%)	32.26	43.21

V. CONCLUSION

The paper proposes a new, computationally efficient and completely unsupervised method for object localization. We have tested our algorithm on challenging and difficult datasets like PASCAL VOC 2007 and Object Discovery. We have achieved comparable results to other weakly supervised learning algorithms that have been proposed for the same problem. Furthermore, we are planning to advance in the

same direction and extend our algorithm to handle multiple object instances per image, which is a common scenario.

REFERENCES

- [1] S. An, P. Peursum, W. Liu, and S. Venkatesh. Efficient algorithms for subwindow search in object detection and localization. In *Proceedings of the IEEE CVPR*, pages 264–271. IEEE, 2009.
- [2] S. Andrews, I. Tsochantaris, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.
- [3] M. Blaschko and C. Lampert. Object localization with global and local context kernels. 2009.
- [4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.
- [5] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *Proceedings of the IEEE CVPR*, pages 2027–2034, 2014.
- [6] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE CVPR*, pages 3286–3293, 2014.
- [7] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE CVPR*, pages 1201–1210, 2015.
- [8] M. Cho, Y. M. Shin, and K. M. Lee. Unsupervised detection and segmentation of identical objects. In *Proceedings of the IEEE CVPR*, pages 1617–1624. IEEE, 2010.
- [9] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- [10] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017.
- [11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE CVPR*, volume 1, pages 886–893. IEEE, 2005.
- [12] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010.
- [13] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.
- [14] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2007 (voc 2007) results (2007), 2008.
- [15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.
- [17] V. Ferrari, L. Fevrier, C. Schmid, and F. Jurie. Groups of adjacent contour segments for object detection. 2008.
- [18] R. Gokberk Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *Proceedings of the IEEE CVPR*, pages 2409–2416, 2014.
- [19] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *Proceedings of the IEEE CVPR*, volume 1, pages 19–25. IEEE, 2006.
- [20] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *Proceedings of the IEEE ICCV*, pages 237–244. IEEE, 2009.
- [21] M. Hoai, L. Torresani, F. De la Torre, and C. Rother. Learning discriminative localization from weakly labeled data. *Pattern Recognition*, 47(3):1523–1534, 2014.
- [22] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE CVPR*, pages 1943–1950. IEEE, 2010.
- [23] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *Proceedings of the IEEE CVPR*, pages 542–549. IEEE, 2012.
- [24] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European Conference on Computer Vision*, pages 253–268. Springer, 2014.
- [25] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Advances in neural information processing systems*, pages 961–969, 2009.
- [26] D. Kuettel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagnet. In *ECCV*, pages 459–473. Springer, 2012.
- [27] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Proceedings of the IEEE CVPR*, pages 1139–1146, 2013.
- [28] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *Proceedings of the IEEE ICCV*, pages 1925–1932. IEEE, 2009.
- [29] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.
- [30] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. 2011.
- [31] C. Rother, T. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *Proceedings of the IEEE CVPR*, volume 1, pages 993–1000. IEEE, 2006.
- [32] H. Rowley and S. Baluja. T. k anade, human face detection in visual scenes. *Carnegie-Mellon University*, 199(6):1, 1995.
- [33] H. A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. In *Advances in Neural Information Processing Systems*, pages 875–881, 1996.
- [34] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE CVPR*, pages 1939–1946, 2013.
- [35] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *Proceedings of the IEEE ICCV*, pages 2984–2991, 2013.
- [36] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *ECCV*, pages 594–608, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- [37] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, pages 594–608. Springer, 2012.
- [38] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *Proceedings of the IEEE ICCV*, pages 343–350. IEEE, 2011.
- [39] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. 2005.
- [40] W. Tahir, A. Majeed, and T. Rehman. Indoor/outdoor image classification using gist image features and neural network classifiers. In *High-Capacity Optical Networks and Enabling/Emerging Technologies (HONET), 2015 12th International Conference on*, pages 1–5. IEEE, 2015.
- [41] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *Proceedings of the IEEE CVPR*, pages 1464–1471, 2014.
- [42] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.
- [43] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *Proceedings of the IEEE CVPR*, pages 2217–2224. IEEE, 2011.
- [44] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In *Proceedings of the IEEE CVPR*, pages 1–8. IEEE, 2008.
- [45] A. Vora and S. Raman. Iterative spectral clustering for unsupervised object localization. *Pattern Recognition Letters*, 106:27 – 32, 2018.
- [46] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, pages 431–445. Springer, 2014.
- [47] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405. Springer, 2014.

Development of an Efficient Low-complexity Channel Estimator for Digital Television Terrestrial Broadcasting Systems

Ghanshyamkumar Sah

Department of Electronics and Communication Engineering
Indian Institute of Technology, Roorkee

Roorkee, India

ghanshyam.shah@gmail.com

Pyari Mohan Pradhan

Department of Electronics and Communication Engineering
Indian Institute of Technology, Roorkee

Roorkee, India

pmpradhan.fec@iitr.ac.in

Abstract—Orthogonal frequency division multiplexing (OFDM) is widely used to transmit data in many wireless communication applications including digital television terrestrial broadcasting (DTTB). Although the existing dual pseudo noise padding (DPNP) based time-domain synchronous OFDM (TDS-OFDM) system has low complexity, the spectral efficiency is low. The time-frequency-domain (TFD) based frame structure enhances the system performance of TDS-OFDM over fast time-varying channels by compromising with computational complexity. This paper proposes a novel frame structure for OFDM-based DTTB system which incorporates pilots in the time domain, and retains the cyclic prefix and modulable orthogonal sequence (MOS) from the TFD-based frame structure. Since the proposed frame structure is completely defined in time domain, channel estimation and equalization become easier. Using the new frame structure, a novel channel estimation technique is proposed that works in two stages. In the first stage, the MOS in guard interval is used to estimate the channel delay and gain. In the second stage, the channel gains estimated in first stage are fine-tuned using adaptive algorithms such as least mean square (LMS) or recursive least squares algorithm. The bit-error-rate (BER) performance of the proposed two-stage channel estimation technique is better compared to that of DPNP-based TDS-OFDM. In addition, computational complexity of the proposed LMS-based two-stage channel estimation approach is low compared to TFD-based TDS-OFDM system. Less than 1.5% of the total sub-carriers are used as redundant pilots, and therefore the loss in spectral efficiency is negligible in the proposed approach.

Index Terms—DTTB, OFDM, Channel estimation

I. INTRODUCTION

Orthogonal frequency division multiplexing (OFDM) is widely used in digital television terrestrial broadcasting (DTTB) over frequency selective multipath channels. In general, the OFDM block transmission schemes can be broadly categorized based on guard interval and pilot placement in the OFDM block. In cyclic prefix OFDM (CP-OFDM) [1] frame structure, the inter-carrier-interference (ICI) and inter-block-interference (IBI) [2], [3] are removed by inserting the guard band, called cyclic prefix, between two OFDM frames. For channel estimation, the pilots are placed in OFDM frame,

which leads to lower spectral efficiency. In zero padding OFDM (ZP-OFDM) [4], zeros are padded instead of CP, to address the problem of channel null [4]. In CP-OFDM and ZP-OFDM, 10% of the subcarriers are used as pilots, which leads to low spectral efficiency. In time-domain synchronous OFDM (TDS-OFDM) [5]–[7], pseudorandom noise (PN) sequence is used as guard band between two OFDM data frames, so as to increase the spectral efficiency. This PN sequence is also exploited for estimation of channel [5] as well as for synchronization. The TDS-OFDM has advantages in terms of spectral efficiency and frame synchronization, but it suffers from the problem of IBI in the training sequence (TS) which consists of PN sequence and data block. The IBI is removed by iterative padding subtraction (IPS) [8] that includes simultaneous estimation and equalization of channel. However, the performance of TDS-OFDM degrades due to IPS.

In order to solve the problem of interference in TDS-OFDM, one of the solutions suggested in the literature is the dual PN padding (DPNP) frame structure [9]. DPNP-based channel estimation is simple and reliable, and has low computational complexity. Thus, it has become the strongest contender among above frame structures for the up-coming DTTB standard [5], [10]. The presence of two PN sequences in DPNP-based TDS-OFDM leads to low spectral efficiency as compared to IPS-based TDS-OFDM. Further, this scheme assumes that the channel is time invariant during each symbol interval, and hence results in high bit-error-rate (BER) in fast fading channels.

For fast time varying fading channels, a time-frequency-domain (TFD) based frame structure is proposed by Linglong Dai *et al.* [11] to improve the performance of TDS-OFDM system. The BER performance of this technique is better than IPS-based TDS-OFDM as well as DPNP-based TDS-OFDM. In this frame structure, path delay and path gain are computed using time-domain TS and frequency-domain pilots, respectively. However, the TFD-based TDS-OFDM has very high computational complexity.

In order to reduce the computational complexity while retaining BER performance and spectral efficiency, this paper

proposes a novel frame structure for TDS-OFDM as well as a low-complexity channel estimator. The major contributions of this paper are listed below:

- The proposed frame structure contains time-domain pilots in the OFDM data block. Thus, it enhances the BER performance in comparison to DPNP-based TDS-OFDM frame structure [9]. Since the proposed frame structure is completely defined in time domain, channel estimation and equalization become easier.
- Two stage channel estimation technique has been proposed for TDS-OFDM systems deployed over fast fading channels. The BER performance and spectral efficiency of proposed approaches are similar to the counterparts based on the TFD-based frame structure [11].
- Two variants of channel estimator are proposed for TDS-OFDM systems based on two widely popular adaptive algorithms, least mean square (LMS) and recursive least square (RLS). The computational complexity of the proposed TS-LMS-based method is very low, and hence it is best suited for applications where BER performance and computation complexity are priorities.

II. SYSTEM MODEL

A. Proposed Frame Structure for OFDM System

In order to increase the DTTB system performance in terms of computational complexity and BER, a signal frame structure for TDS-OFDM is proposed. Fig. 1 shows its comparison with the frame structures of DPNP-based and TFD-based TDS-OFDM systems. The proposed frame structure is different from the DPNP-based scheme in terms of time-domain guard band and pilots in OFDM block. It is also different from the TFD-based scheme in terms of the pilots in OFDM block. In proposed method, the pilots are placed in time domain, while in TFD-based scheme, the pilots are in frequency domain.



Fig. 1. Proposed TDS-OFDM frame structure

B. Modelling of Guard Interval and OFDM Data Block

In the proposed TDS-OFDM frame structure, the guard interval structure proposed in [11] is followed. The i^{th} symbol of the TDS-OFDM data block,

$$s_i = [s_{i,0}, s_{i,1}, \dots, s_{i,P-1}]^T$$

consists of guard interval known a priori,

$$g_i = [c_{i,1}, \dots, c_{i,M-1}, c_{i,0}, c_{i,1}, \dots, c_{i,M-1}]^T$$

which is of length $2M - 1$, and the data block

$$x_i = [x_{i,0}, x_{i,1}, \dots, x_{i,N-1}]^T$$

of length N . The guard interval g_i comprises of TS $c_i = [c_{i,0}, c_{i,1}, \dots, c_{i,M-1}]^T$ having length M and its corresponding cyclic prefix $[c_{i,1}, \dots, c_{i,M-1}]^T$ with length $M - 1$. In the time domain, data block is expressed by $x_i = F_N^H X_i$, where $X_i = [X_{i,0}, X_{i,1}, \dots, X_{i,N-1}]^T$ denotes the fast Fourier transform (FFT) of x_i . $P = N + 2M - 1$ represents the complete duration of a TDS-OFDM symbol.

After shifting the TS c_i by one sample cyclically to the right for the i^{th} symbol, new TS c_{i+1} for the $(i + 1)^{\text{th}}$ symbol is obtained as

$$c_{i+1} = \begin{bmatrix} 0_{1 \times (M-1)} & 1 \\ I_{M-1} & 0_{(M-1) \times 1} \end{bmatrix} c_i. \quad (1)$$

The TS is not constant in this scheme. It is shown in [12] that for channel estimation, constant TS is not a good choice. From (1), the new guard interval for $(i + 1)^{\text{th}}$ TDS-OFDM symbol can be obtained as

$$g_{i+1} = [c_{i,0}, c_{i,1}, \dots, c_{i,M-2}, c_{i,M-1}, c_{i,0}, c_{i,1}, \dots, c_{i,M-2}]^T$$

The first M samples of g_{i+1} is same as the last M samples of g_i . The PN sequence of DPNP-based TDS-OFDM does not have ideal autocorrelation property, and hence not best suited for performing channel estimation [13], [14]. Therefore, the modulable orthogonal sequence (MOS) [11] is used as TS, i.e.

$$c_{i,n} = b(n_1) \exp\left(\frac{2\pi}{\sqrt{M}} mn_0 n_1\right), \quad 0 \leq n \leq M - 1 \quad (2)$$

where $0 \leq n_0 \leq \sqrt{M} - 1$, $0 \leq n_1 \leq \sqrt{M} - 1$, $n = n_0\sqrt{M} + n_1$, m is relatively prime to \sqrt{M} , and $|b(n_1)| = 1$. In this study, $m = 1$ and $b(n_1) = 1$ are adopted for ease. The MOS has precise autocorrelation property that can be represented by

$$c_i \otimes c_i = M [1 \ 0_{1 \times (M-1)}]^T \quad (3)$$

In the DPNP-based TDS-OFDM shown in Fig. 1(a), all data subcarriers carry data, and no pilot is embedded in the data block [8]–[10]. In the proposed frame structure, N_d data subcarriers and N_p pilots are used in the data block of OFDM. The TFD-based TDS-OFDM includes pilots in frequency domain leading to higher complexity [11]. Therefore, the proposed frame structure includes pilots in time domain. $N = N_d + N_p$, and N_p is very small compared to N . Thus, the spectral loss in the proposed frame structure is low.

C. System Model Over Fading Channels

As there is one sample shifted TS in (1), the data block can be easily reconstructed from the received OFDM frame. However, this received OFDM frame suffers from IBI. By using the method of add-subtract [9], the data block without IBI can be reconstructed. At the receiver, the data block $y_i = [y_{i,0}, y_{i,1}, \dots, y_{i,N-1}]^T$ in time domain is defined as

$$y_{i,n} = \sum_{l=0}^{L-1} h_{i,n,l} x_{i,(n-n_l)_N} + w_{i,n} \quad (4)$$

where, $w_{i,n}$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ^2 . $h_{i,n,l}$ represents the path gain.

Sub-script l indicates the l^{th} path, n is the time instant, and i represents the OFDM block. These path gains are non-zero after delay of n_l ($n_0 = 0$ is assumed in this paper). L is the total number of multi-paths. The data block and guard band can have interference if maximum channel length n_{L-1} is larger than the TS length M . To avoid this, $n_{L-1} < M$. The total number of multi-paths is very very small compared to the maximum channel length, i.e., $L \ll n_{L-1}$ [15], [16].

After taking FFT of the received sequence, the frequency-domain representation of the data block $Y_i = [Y_{i,0}, Y_{i,1}, \dots, Y_{i,N-1}]^T$ is [3]

$$\begin{aligned} Y_{i,k} &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} y_{i,n} e^{-j \frac{2\pi}{N} nk} \\ &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} \left(\sum_{l=0}^{L-1} h_{i,n,l} x_{i,(n-n_l)} + w_{i,n} \right) e^{-j \frac{2\pi}{N} nk} \quad (5) \\ &= X_{i,k} H_{i,k,k} + \underbrace{\sum_{q=0, q \neq k}^{N-1} X_{i,q} H_{i,k,q}}_{\text{ICI}} + W_{i,k} \end{aligned}$$

where $W_{i,k} = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} w_{i,n} e^{-j \frac{2\pi}{N} nk}$ is the noise term, and

$$H_{i,k,q} = \frac{1}{N} \sum_{l=0}^{L-1} \left(\sum_{n=0}^{N-1} h_{i,n,l} e^{-j \frac{2\pi}{N} n(k-q)} \right) e^{-j \frac{2\pi}{N} q n_l}. \quad (6)$$

The typical TDS-OFDM schemes [8]–[10] assume that the channel is time-invariant within each symbol duration, i.e., $h_{i,0,l} = h_{i,1,l} = \dots = h_{i,N-1,l} = h_{i,l}$ ($0 \leq l \leq L-1$), where $h_{i,l}$ represents the average path gain for l^{th} path, i.e., $h_{i,l} = \frac{1}{N} \sum_{n=0}^{N-1} h_{i,n,l}$. Then, the ICI term $H_{i,k,q}$ ($q \neq k$) becomes zero, i.e., $H_{i,k,q} = 0$, and (5) is consequently simplified as

$$Y_{i,k} = X_{i,k} H_{i,k,k} + W_{i,k}. \quad (7)$$

where $H_{i,k} = H_{i,k,k}$. The system model presented by [17] assumes that the channel is time-invariant within each TDS-OFDM symbol. In practical scenario, the channel is time-variant [3], and therefore there will be performance loss in terms of BER. This problem is solved in [11].

III. PROPOSED TWO-STAGE CHANNEL ESTIMATION

In DPNP-based TDS-OFDM, the complete estimation of channel is carried out using the TS present in the guard band. In TFD-based TDS-OFDM, path delay and path gain estimation are carried out using TS in time domain and frequency-domain pilots in data block, respectively. Unlike these two methods, in the proposed technique, both path delay and gain are estimated by using TS in guard interval, and further the path gains are fine-tuned by using pilots in data block in time domain.

A. Channel Impulse Response (CIR) Estimation using TS

The guard interval of the proposed structure comprises of TS and a cyclic extension of CP. Due to the CP in the guard interval, the received TS $d_i = [d_{i,0}, d_{i,1}, \dots, d_{i,M-1}]^T$

is susceptible to the IBI between the guard interval and data block. Due to cyclic extension of CP in the guard interval, the received d_i is defined as

$$d_i = c_i^{(s)} \otimes h_i + v_i \quad (8)$$

Using the relationship between the circular convolution and circular correlation, $c_i^{(s)}$ is obtained from c_i , where $c_i^{(s)} = [c_{i,0}, c_{i,M-1}, c_{i,M-2}, \dots, c_{i,1}]^T$. $v_i = [v_{i,0}, v_{i,1}, \dots, v_{i,M-1}]^T$ represents the AWGN vector with zero mean and variance σ^2 , and

$$h_i = [h_{i,0}, \underbrace{0, \dots, 0}_{n_1 - n_0 - 1}, h_{i,1}, 0, \dots, 0, h_{i,L-1}, \underbrace{0, \dots, 0}_{M - n_{L-1} - 1}]^T$$

denotes the $M \times 1$ zero padded CIR.

The rough estimate of channel \hat{h}_i is obtained by the circular correlation between $c_i^{(s)}$, which is known at the receiver and the received TS d_i as

$$\begin{aligned} \hat{h}_i &= \frac{1}{M} c_i^{(s)} \otimes d_i = \frac{1}{M} c_i^{(s)} \otimes (c_i^{(s)} \otimes h_i + v_i) \\ &= h_i + \frac{1}{M} c_i^{(s)} \otimes v_i \end{aligned} \quad (9)$$

Eq. (9) uses the property of perfect autocorrelation of the MOS mentioned in (3).

B. Fine-Tuning of CIR Estimate using Pilots

The rough channel gains estimated by using TS of guard interval, are fine-tuned by using pilots in data block in time domain. The conventional LMS/RLS algorithm based channel estimation is used for the second stage.

For updating CIR coefficients using LMS/RLS algorithm, initial weights are taken as the previously estimated coefficients in (9). The pilots are inserted in time domain in the data block as $p_{i,n} = \{p_{i,0}, p_{i,1}, \dots, p_{N_p-1}\}$. N_p should be greater than maximum channel delay n_L , i.e. $N_p \geq n_L$. The desired output $d_{i,n}$ is given by

$$d_{i,n} = h_{i,n}^H p_{i,n} + w_{i,n} \quad (10)$$

where $h_{i,n}$ are the actual channel coefficients. When the pilots $p_{i,n}$ are passed through the filter $\hat{h}_{i,n}$, the output $o_{i,n}$ is given by

$$o_{i,n} = \hat{h}_{i,n}^H p_{i,n} \quad (11)$$

Thus, the estimation error or the error signal $e_{i,n}$ is defined as

$$e_{i,n} = d_{i,n} - o_{i,n} \quad (12)$$

The tap-weight adaptation using the LMS algorithm is represented as

$$\hat{h}_{i,n+1} = \hat{h}_{i,n} + \mu p_{i,n} e_{i,n} \quad (13)$$

where μ is step-size parameter whose range is defined as

$$0 < \mu < \frac{2}{N_f S_{\max}}$$

S_{\max} is the maximum value of the power spectral density of the input $d_{i,n}$, and N_f represents the filter length.

Similarly, the tap-weight adaptation using RLS algorithm can be defined as

$$\hat{h}_{i,n+1} = \hat{h}_{i,n} + k_{i,n} e_{i,n} \quad (14)$$

where $k_{i,n}$ is the gain vector defined as

$$k_{i,n} = \frac{L_{i,n-1} p_{i,n}}{\lambda + p_{i,n}^H L_{i,n-1} p_{i,n}}$$

$L_{i,n}$ is the inverse correlation matrix represented as

$$L_{i,n} = \lambda^{-1} L_{i,n-1} - \lambda^{-1} k_{i,n} p_{i,n}^H L_{i,n-1}$$

λ is the exponential weighting factor or forgetting factor which is chosen in the range $0 < \lambda \leq 1$.

IV. SIMULATION RESULTS AND PERFORMANCE ANALYSIS

Simulation study is carried in MATLAB 2016 environment to validate the performance of the designed framework. The simulation parameters used for the study are listed in Table I.

TABLE I
SIMULATION PARAMETER

Parameter	Specification	
Signal bandwidth	7.56 MHz	
Central frequency	770 MHz	
Modulation scheme	QPSK	
OFDM block size (N)	4096	
Guard band size	$2M - 1$; $M = 256$	
Resolvable path (L)	6	
Doppler spread (f_d)	10 Hz and 100 Hz	
Total pilots (N_p)	60	
Parameters of the Brazil D multipath channel		
Path Number	Delay (μs)	Attenuation (dB)
1	0.15	0.1
2	0.63	3.8
3	2.22	2.6
4	3.05	1.3
5	5.86	0
6	5.93	2.8

A. Path Delay Estimation

Fig. 2 shows a comparison between channel estimate \hat{h}_i at stage-1 and actual channel h_i at SNR of 10 dB. Due to the perfect auto-correlation property of the MOS, path delay estimation is perfect. However, the channel gain estimates \hat{h}_i deviate from the actual channel h_i .

B. Path Gain Estimation

Fig. 3 shows a comparison between actual channel h_i and the CIR estimated at stage-2 by using proposed LMS/RLS-based method at SNR of 10 dB. It can be observed that channel gain estimates at stage-2 are better than those after stage-1.

Fig. 4 shows the mean square deviation (MSD) or mean square error between actual channel and estimated channel by DPNP-based, proposed TS-LMS-based, proposed TS-RLS-based and TFD-based TDS-OFDM for stationary user. The MSD between actual channel and DPNP-based scheme is

found to be maximum. A considerable reduction in MSD can be noticed at lower SNR, when channel estimates are tuned by proposed LMS/RLS-based method.

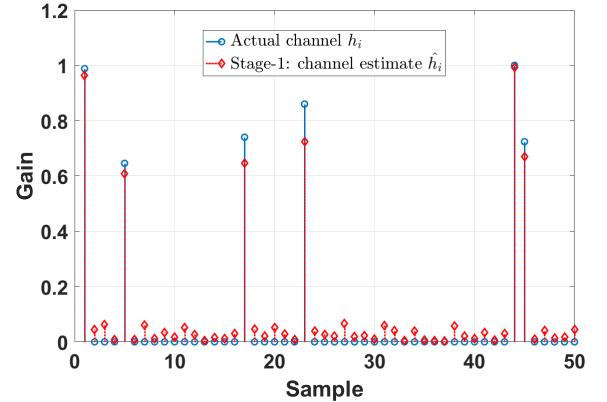


Fig. 2. Path delay estimation at SNR of 10 dB for one instance of channel

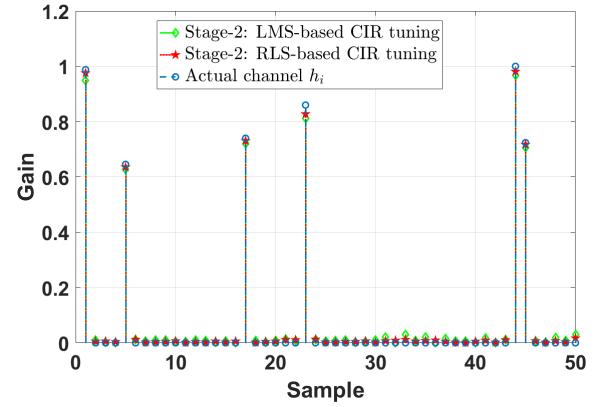


Fig. 3. Comparison of CIR for one instance of channel

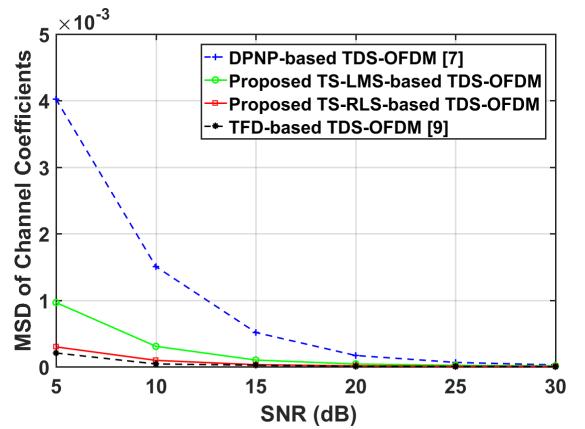


Fig. 4. Mean square deviation of channel coefficients for one instance of channel

C. BER Performance

For BER analysis, the CP-OFDM signal is initially reconstructed from the received TDS-OFDM frame by add-subtract method [9]. Further, the signal can be recovered using any equalization method. The channel equalization using a one-tap zero forcing (ZF) equalizer can be represented as

$$\{\hat{s}_{i,n}\}_{n=0}^{N-1} = \frac{FFT_N(\{y_{i,n}\}_{n=0}^{N-1})}{FFT_N(\{\hat{h}_{i,n}\}_{n=0}^{L-1})} \quad (15)$$

Fig. 5 shows the BER performance of different TDS-OFDM systems for a stationary user over the Brazil D channel. The proposed techniques show the improvement in BER performance in comparison to DPNP-based scheme. Additionally, it can be observed that the performance of TFD-based scheme is moderately better than the proposed methods. The moderate edge in BER performance of the TFD-based scheme is due to independent channel path gain estimation, using the inserted pilots in the frequency domain in the data block. However, this advantage in BER performance is at the cost of huge computational complexity compared to the proposed LMS-based method. RLS-based method also shows better BER performance than LMS-based method, but at the cost of high computational complexity.

Fig. 6 shows the BER performance of different TDS-OFDM systems for a user moving with a speed of 14 km/h. The performance has slightly degraded due to time varying channel, which can also be observed by comparing Figs. 5 and 6. At the speed of 14 km/h and Doppler spread of $f_d = 10$ Hz, the BER performances of proposed TS-LMS-based scheme and TS-RLS-based scheme are better than DPNP-based scheme. The difference in the BER of the proposed scheme and the TFD-based scheme have reduced due to better convergence of LMS/RLS algorithm.

Fig. 7 shows the BER performance of different TDS-OFDM systems for a user moving with a speed of 140 km/h in a Brazil D channel. The overall BER performance has slightly degraded as the speed has increased. At a speed of 140 km/h

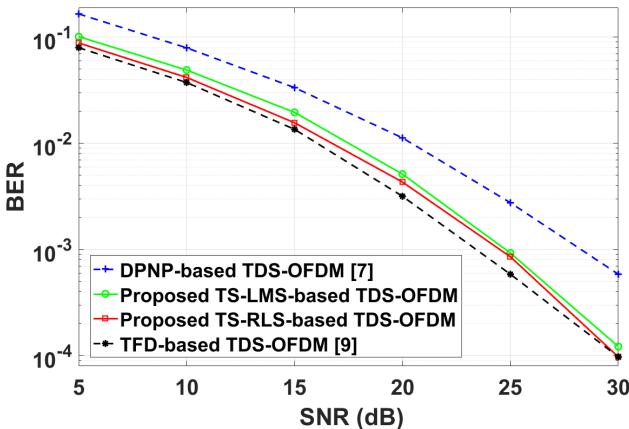


Fig. 5. BER performance comparison for stationary user over the Brazil D channel

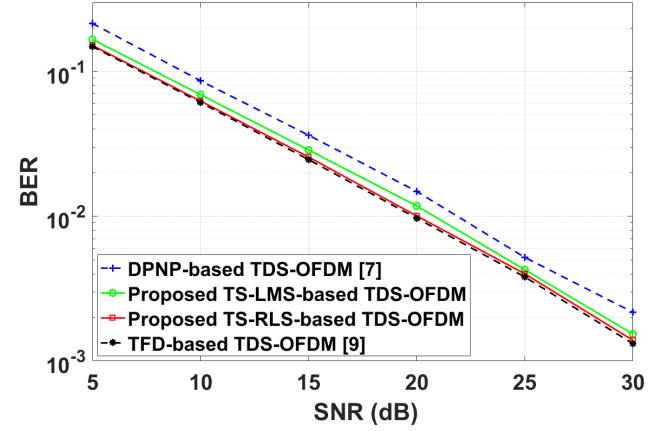


Fig. 6. BER performance comparison at the speed of 14 km/h over the Brazil D channel

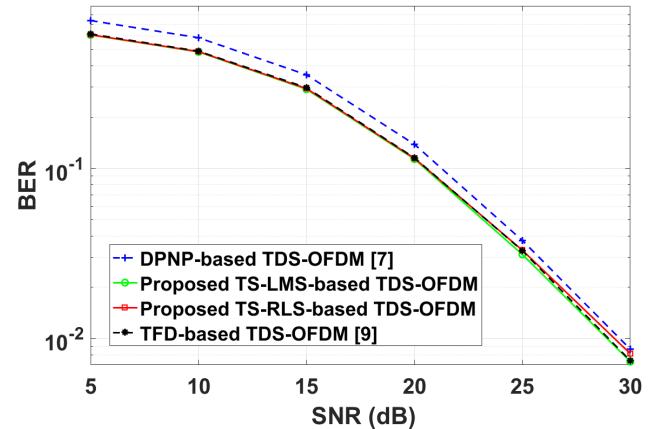


Fig. 7. BER performance comparison at the speed of 140 km/h over the Brazil D channel

and Doppler spread of $f_d = 100$ Hz, the BER performances of proposed TS-LMS-based and TS-RLS-based schemes are better than DPNP-based scheme, and similar to TFD-based scheme.

D. Spectral Efficiency

In the proposed frame structure, N_p pilots are inserted in data block. The CIR is fine-tuned by using the LMS/RLS algorithm. For proper convergence of LMS/RLS algorithm, channel length n_L should be smaller than inserted pilot N_p . These pilots are redundant in nature, and do not contain any data information, leading to loss in spectral efficiency. Thus, the loss in spectral efficiency for the proposed method is given by

$$E_{\text{loss}} = \frac{N}{N+2M} - \frac{N-N_p}{N+2M-1} \quad (16)$$

Table I shows that pilots occupy approx. 1.46% of total subcarriers in the proposed frame structure, and hence the loss in the spectral efficiency is insignificant. Summary of spectral efficiency of the frame structures is provided in Table II.

TABLE II
COMPARISON OF SPECTRAL EFFICIENCY

TDS-OFDM	Spectral Efficiency
DPNP-based	88.88%
TFD-based	88.13%
Proposed TS-LMS-based	87.60%
Proposed TS-RLS-based	87.60%

TABLE III
COMPARISON OF NUMBER OF COMPLEX MULTIPLICATIONS

TDS-OFDM	Generalized Computational Complexity	Case Study
DPNP-based	$M \log_2 M$	2,048
TFD-based	$M \log_2 M + 2N_{group}(Q+1)^2 L^2 + (Q+1)^3 L^3 + N_{group}(Q+1)L + 2L(2d+1)N$	1,54,832
Proposed TS-LMS-based	$M \log_2 M + N_p(2N_p + 3)$	9,428
Proposed TS-RLS-based	$M \log_2 M + N_p(3N_p^2 + 11N_p + 8)$	6,88,080

E. Computational Complexity

In this study, the number of complex multiplications is taken as a metric to compare the computational complexity of different schemes. The CIR estimation in (9) requires a M -point FFT as well as a M -point inverse FFT (IFFT). Each M -point FFT/IFFT needs $(M/2)\log_2 M$ complex multiplications. This computation is performed for all frame structures discussed in this paper. In addition, for tuning the CIR by proposed LMS-based method, $N_p(2N_p + 3)$ multiplications are needed. For tuning the CIR by proposed RLS-based method, $N_p(3N_p^2 + 11N_p + 8)$ multiplications are needed. The TFD-based TDS-OFDM requires $2N_{group}(Q+1)^2 L^2 + (Q+1)^3 L^3$ multiplications for computation of the Moore-Penrose inverse matrix, and $N_{group}(Q+1)L$ complex multiplications for matrix multiplication [11]. $2N_{group}$ represents the number of grouped pilots. The maximum Doppler spread of the channel is used to decide the factor Q . In addition, the TFD-based TDS-OFDM needs $2L(2d+1)N$ multiplications to calculate the ICI coefficients [11]. Using the parameters listed in Table I, the number of complex multiplications required by different schemes are summarized in Table III.

V. CONCLUSIONS

This paper proposes a novel low complexity two stage channel estimation scheme based on a novel OFDM frame structure which enhances the performance of TDS-OFDM systems. This scheme employs the TS (one-sample shifted) in the guard interval to estimate the CIR. Further, the pilots inserted in the data block is used to fine-tune the CIR by using LMS/RLS algorithm. The proposed scheme leads to better channel estimation for stationary, slow and fast moving users. The proposed scheme could accurately pursue the variations in channels, and outperforms the DPNP-based TDS-OFDM methods in terms of BER. The overall BER performance degradation is due to the ICI becoming significant at higher velocity. The proposed method slightly under performs in

terms of BER for stationary and slow moving user when compared to TFD-based TDS-OFDM. Computational complexity of the proposed TS-LMS-based method is low whereas that of proposed TS-RLS-based scheme is high compared to TFD-based scheme. Thus, the proposed TS-LMS-based TDS-OFDM is best suited if BER performance and computation complexity are priorities.

REFERENCES

- [1] Y. S. Cho, J. Kim, W. Y. Yang, and C. G. Kang, *MIMO-OFDM Wireless Communications with MATLAB*. Wiley Publishing, 2010.
- [2] H. Wu, X. Huang, and D. Xu, "Novel semi-blind ICI equalization algorithm for wireless OFDM systems," *IEEE Transactions on Broadcasting*, vol. 52, no. 2, pp. 211–218, Jun. 2006.
- [3] H. Wu, X. Huang, Y. Wu, and X. Wang, "Theoretical studies and efficient algorithm of semi-blind ICI equalization for OFDM," *IEEE Transactions on Wireless Communications*, vol. 7, no. 10, pp. 3791–3798, Oct. 2008.
- [4] B. Muquet, Z. Wang, G. B. Giannakis, M. de Courville, and P. Duhamel, "Cyclic prefixing or zero padding for wireless multicarrier transmissions?" *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 2136–2148, Dec. 2002.
- [5] L. Dai, Z. Wang, and Z. Yang, "Next-generation digital television terrestrial broadcasting systems: Key technologies and research trends," *IEEE Communications Magazine*, vol. 50, no. 6, pp. 150–158, Jun. 2012.
- [6] M. Baaran, H. enol, S. Erkk, and H. A. rpan, "Channel estimation for tds-ofdm systems in rapidly time-varying mobile channels," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8123–8135, Dec. 2018.
- [7] E. Farouk, M. Z. Saleh, M. Ibrahim, and S. Elramly, "Joint channel estimation for tds-ofdm based on superimposed training," in *14th International Conference on Telecommunications (ConTEL)*, June 2017, pp. 55–62.
- [8] J. Wang, Z.-X. Yang, C.-Y. Pan, J. Song, and L. Yang, "Iterative padding subtraction of the PN sequence for the TDS-OFDM over broadcast channels," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 4, pp. 1148–1152, Nov. 2005.
- [9] J. Fu, J. Wang, J. Song, C. Y. Pan, and Z. X. Yang, "A simplified equalization method for dual PN-sequence padding TDS-OFDM systems," *IEEE Transactions on Broadcasting*, vol. 54, no. 4, pp. 825–830, Dec. 2008.
- [10] Z. Yang, L. Dai, J. Wang, J. Wang, and Z. Wang, "Transmit diversity for TDS-OFDM broadcasting system over doubly selective fading channels," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 135–142, Mar. 2011.
- [11] L. Dai, Z. Wang, J. Wang, and Z. Yang, "Joint time-frequency channel estimation for time domain synchronous OFDM systems," *IEEE Transactions on Broadcasting*, vol. 59, no. 1, pp. 168–173, Mar. 2013.
- [12] O. Rousseau, G. Leus, P. Stoica, and M. Moonen, "Gaussian maximum-likelihood channel estimation with short training sequences," *IEEE Transactions on Wireless Communications*, vol. 4, no. 6, pp. 2945–2955, Nov. 2005.
- [13] H. Wu and X. Huang, "Joint phase/amplitude estimation and symbol detection for wireless ICI self-cancellation coded OFDM systems," *IEEE Transactions on Broadcasting*, vol. 50, no. 1, pp. 49–55, Mar. 2004.
- [14] H. Wu and Y. Wu, "Distributive pilot arrangement based on modified m-sequences for OFDM intercarrier interference estimation," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1605–1609, May 2007.
- [15] L. Dai, Z. Wang, and Z. Yang, "Time-frequency training OFDM with high spectral efficiency and reliable performance in high speed environments," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 4, pp. 695–707, May 2012.
- [16] Z. Tang, R. C. Cannizzaro, G. Leus, and P. Banelli, "Pilot-assisted time-varying channel estimation for OFDM systems," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2226–2238, May 2007.
- [17] M. Huemer, A. Onic, and C. Hofbauer, "Classical and bayesian linear data estimators for unique word OFDM," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6073–6085, Dec. 2011.

Power Domain NOMA Design Based on MBER Criterion

Amit Kumar Dutta, *Member, IEEE*,
Indian Institute of Technology Kharagpur, India
Email: amitdutta@gssst.iitkgp.ac.in

Abstract—Non-orthogonal multiple access (NOMA) has been gaining notable attention in the context of next generation communication system. The key primary benefit includes the higher spectrum efficiency compared to its various orthogonal counterparts. In this treatise, we consider a power NOMA design based on the minimum bit error ratio (MBER) criterion. Inspiration has been drawn from the MBER based works, which show a considerable performance improvement in terms of bit error ratio (BER) for a system. In this work, we have considered a single-input single-output (SISO) system with quadrature phase shift keying (QPSK) signal constellation. The numerical results demonstrate an overall BER improvement compared to the existing schemes, albeit, the scheme attracts large computational complexity. Traditionally, NOMA increases the spectral efficiency compared to its orthogonal counterpart. Nevertheless, our proposed solution will still hold this feature along with a better BER performance, though its spectral efficiency will be less compared to the traditional sum-rate based power NOMA.

Index Terms—Power NOMA, MBER, SISO, BER.

I. INTRODUCTION

Effective utilization of physical resources mainly spectrum, time and power has attained notable attention since long time. On the other hand, due to the magnificent tele-traffic growth, multiplexing of limited resources have become inevitable among wireless nodes. One such scheme is the time-division multiple access (TDMA), which has been well researched and deployed in modern wireless technology. Another such scheme is orthogonal frequency division multiple access (OFDMA), which has become a key technological enabler for fourth or fifth generation (4G/5G) cellular standards. However, such orthogonal resource sharing schemes come at the cost of reduced spectrum efficiency.

On the other hand, non-orthogonal multiple access (NOMA) has attracted tremendous interest both from the academia and industry for the next generation wireless protocols, primarily because it allows data transmission for multiple users with the same resource. The key advantage is its effective utilization and sharing of physical resources with the higher spectral efficiency compared to orthogonal multiple access (OMA). There are diversified types of NOMA, namely power domain NOMA, cognitive radio (CR) based NOMA and various others [1], [2]. We focus only on the power NOMA in this paper. Work in [3] shows that the optimal distribution of transmit power among various users can be accomplished by optimizing the sum rate, while [4] shows that power NOMA with fairness constraints outperforms OMA with even the worst channel

condition. A joint bandwidth and power allocation scheme is proposed in [5], [6] with the optimization of sum rate plus allocated bandwidth. Work in [7] deals with power allocation with a certain quality of service (QoS) constraints along with sum rate optimization and max-min fairness. However, a comprehensive survey of power NOMA is available in [8].

The cost function (CF) considered in all the schemes of power NOMA is mainly based on sum rate optimization along with various constraints, which may not optimize the BER performance. In this work, we have conceived minimum bit error ratio (MBER) criterion based CF for this type of NOMA design, which will optimize the BER as well. However, it will not optimize the rate. We will show that our proposed solution is not too far from the rate achieved with the traditional method. The MBER criterion has been conceptualized in [9] in the context of an inter-symbol-interference (ISI) channel system. It is shown in various literature that if the received signal is non-Gaussian, then an SNR gain of 2 – 6 dB can be achieved. Work in [10] designs a precoder based on the MBER criterion with an SNR gain of almost 4 – 6 dB. A relay design in [11] and a cognitive radio (CR) system in [12] have been shown to outperform their other linear counterparts in terms of BER.

Given this background, the contribution of this work is as follows

- 1) We have proposed a power NOMA based on the MBER criterion in this work for an uncoded system. We have considered a SISO system for both the base station (BS) and the users. The data constellation is considered to be from QPSK. The BS will allocate a fraction of the total transmit power to each user and these fractional coefficients of the total power distribution among various users will be determined based on the MBER criterion. The BER will be considered for each user and the proposed method will minimize the overall average BER among all the users.

It is assumed that the BS will get feedback from each user about its instantaneous channel coefficient and the noise variance. Point-to-point channel link between the BS and each user is assumed to be frequency-flat fading one.

Notation: Bold upper and lower case letters denote matrix and vector, respectively. The superscripts $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate of a matrix, respectively. $\mathbb{E}[\cdot]$ denotes the expectation, while \mathbf{I}_N denotes $(N \times N)$ -element

identity matrix. For any complex number x , $\Re(x)$ and $\Im(x)$ represent its real and imaginary parts, respectively.

II. SYSTEM MODEL

We consider a BS with N number of users connected to it as shown in Fig. 1. Each user and the BS is assumed to have

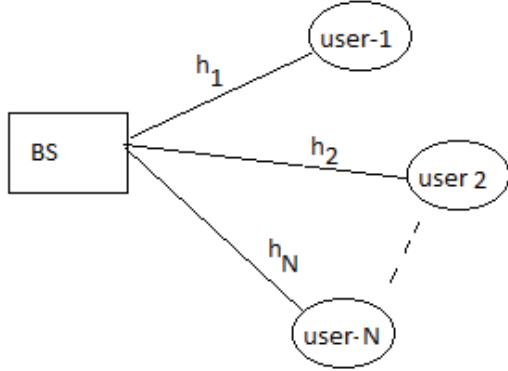


Fig. 1. A typical power NOMA scheme with N users attached to a base station.

a single antenna and the point-to-point channel is assumed to be frequency flat fading. Let us assume that the channel coefficient between BS and the i^{th} user is h_i . Let us also assume that the total power to be transmitted from the BS is P_T . BS will allocate $\alpha_i P_T$ power to the i^{th} user, where $0 < \alpha_i < 1$ indicates the fractional part of P_T corresponding to this user. Let us assume that x_i is the intended data to be transmitted to the i^{th} user. The composite transmitted data at the BS is written as

$$x = \sum_{i=1}^N \sqrt{\alpha_i P_T} x_i. \quad (1)$$

Subsequently, the received signal y_i at the i^{th} user can be written as follows

$$y_i = h_i \left[\sum_{i=1}^N \sqrt{\alpha_i P_T} x_i \right] + w_i, \quad (2)$$

where w_i is the zero-mean additive white Gaussian noise (AWGN) with variance $\sigma_{w,i}^2$. It is assumed that w_i is independent across all the users. We also assume that users are ordered as per their instantaneous channel gain and noise variance ratio, i.e. $\frac{|h_1|^2}{\sigma_{w,1}^2} > \frac{|h_2|^2}{\sigma_{w,2}^2} > \dots > \frac{|h_N|^2}{\sigma_{w,N}^2} > 0$.

For this power domain NOMA configuration, BS uses "superposition" method to transmit all users' data, while each user deploys successive-interference-cancellation (SIC) [13] method to cancel out data from users below its order for decoding its own data. Hence, at the i^{th} ordered user, the user will first decode data from $i+1, i+2, \dots, N^{th}$ users and these are discarded from the y_i , while data from $1, 2, \dots, (i-1)^{th}$ users will be taken as interference.

III. COST FUNCTION

The cost function (CF) involves the probability of bit error at each receiver. Painstakingly, dealing with exact expression for BER is tougher compared to the symbol error ratio (SER). However, for the uncoded system, the BER can be linearly approximated with the SER as $BER \approx \frac{SER}{\log_2 M}$ at higher SNR [14]. In literature, the MBER problems have dealt with the SER mainly [10], [12], considering this slight abuse of notion. Hence, in this context, SER and BER will be used interchangeably.

The SER is calculated for each user and it depends on the exact constellation being considered. It is evident that the SER depends on the variable of interest α_i . Assume that the SER at the i^{th} user is defined as $P_e^i(\alpha)$, where $\alpha = [\alpha_1 \alpha_2 \dots \alpha_N]^T$. Hence, the composite average CF is defined as follows

$$P_e(\alpha) = \frac{1}{N} \sum_{i=1}^N P_e^i(\alpha). \quad (3)$$

We will also ensure that each user maintains a minimum rate, say R_i^{req} , as is done for the traditional sum rate based power NOMA design. Assuming that the rate for the i^{th} user is R_i , the proposed optimization problem can be formulated as follows

$$\begin{aligned} \alpha^{opt} &= \arg \min_{\alpha} P_e(\alpha), \\ s.t \quad (1) \quad &\sum_{i=1}^N \alpha_i = 1, \\ (2) \quad &\alpha_i > 0 \text{ for } i = 1, 2, \dots, N, \\ (3) \quad &R_i \geq R_i^{req} \text{ for } i = 1, 2, \dots, N, \end{aligned} \quad (4)$$

where α^{opt} is the optimized value of this parameter vector. The R_i can be formulated as [3]

$$R_i = \log_2 \left(1 + \frac{P_t \sigma_x |h_i|^2 \alpha_i}{\sigma_{w,i}^2 + P_t \sigma_x^2 |h_i|^2 \sum_{j=1}^{i-1} \alpha_j} \right), \quad (5)$$

where σ_x^2 is the average constellation power.

A. CF Development

The CF at each user is specific to the choice of signal constellation. In this work, we choose QPSK as proof of concept in the context of power NOMA development based on MBER criterion. Let us assume that the input signal set x_i is from QPSK constellation set. Let us assume that the i^{th} user will decode and eliminate the interference perfectly using the SIC method for $i+1 \leq j \leq N$ users. Hence, the post-SIC data at the i^{th} user is written as follows

$$r_i = x_i + \frac{w_i}{h_i \sqrt{\alpha_i P_T}} + \sum_{j=1}^{i-1} \sqrt{\frac{\alpha_j}{\alpha_i}} x_j. \quad (6)$$

The constellation is chosen as $x_i \in \frac{\sigma_x}{\sqrt{2}}(\pm 1 \pm 1j)$. Hence, the probability of correct symbol decision i.e $P_c^i(\alpha)$ for the i^{th} user will be as follows

$$P_c^i(\alpha) = \Pr[\Re(r_i)\Re(x_i) > 0] \times \Pr[\Im(r_i)\Im(x_i) > 0]. \quad (7)$$

It is noted that

$$P_e^i(\alpha) = 1 - P_c^i(\alpha). \quad (8)$$

Assume that $\sqrt{\alpha_i} \triangleq [\sqrt{\alpha_1} \sqrt{\alpha_2} \dots \sqrt{\alpha_i}]^T$ and $\mathbf{x}^i \triangleq [x_1 x_2 \dots x_i]$. We express $P_c^{i,R}(\alpha) \triangleq \Pr[\Re(r_i)\Re(x_i) < 0]$ in (9). Similarly, we express $P_c^{i,I}(\alpha) \triangleq \Pr[\Im(r_i)\Im(x_i) < 0]$ in (9). For a QPSK data set, as $\Re(x_i)$ or $\Im(x_i)$ represent only $\pm \sigma_x/\sqrt{2}$, it is evident that $L_i = 2^i$. Let us define $z_{i,j}^R$ as follows

$$\begin{aligned} z_{i,j}^R &\triangleq \Re\left[x_i + \frac{w_i}{h_i\sqrt{\alpha_i P_T}} + \frac{1}{\sqrt{\alpha_i}}\sqrt{\alpha_{i-1}}^T \mathbf{x}^{j-1}\right] \Re(x_i), \\ &\triangleq \frac{\sigma_x^2}{2} + \hat{w}_i^R + b_{i,j}^R, \\ &\triangleq c_{i,j}^R + \hat{w}_i^R, \end{aligned} \quad (11)$$

where $\hat{w}_i^R \triangleq \Re\left[\frac{w_i}{h_i\sqrt{\alpha_i P_T}}\right] \Re(x_i)$ along with $b_{i,j}^R \triangleq \frac{1}{\sqrt{\alpha_i}}\sqrt{\alpha_{i-1}}^T \Re[\mathbf{x}^{i-1}(j)] \Re(x_i)$ and finally $c_{i,j}^R \triangleq \frac{\sigma_x^2}{2} + b_{i,j}^R$. Hence, \hat{w}_i is a zero-mean Gaussian random variable with variance $\sigma_{\hat{w},i}^2 \triangleq \frac{\sigma_w^2 \sigma_x^2}{4|h_i|^2 \alpha_i P_T}$. Therefore, $P_c^{i,R}(\alpha)$ would become

$$\begin{aligned} P_c^{i,R}(\alpha) &= \frac{1}{L_i} \sum_{j=1}^{L_i} \Pr[c_{i,j}^R + \hat{w}_i^R > 0 | \mathbf{x}^i(j)], \\ &= \frac{1}{L_i} \sum_{j=1}^{L_i} \left[1 - Q\left(\frac{c_{i,j}^R}{\sigma_{\hat{w},i}^2}\right) \right], \end{aligned} \quad (12)$$

where $Q(x)$ is the Gaussian error function defined as $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{x^2}{2}) dx$ [14]. The above equation represents the probability of correct decision, considering all possible realizations of $\mathbf{x}^i(j)$. Similarly, $P_c^{i,I}(\alpha)$ can be expressed as follows

$$P_c^{i,I}(\alpha) = \frac{1}{L_i} \sum_{j=1}^{L_i} \left[1 - Q\left(\frac{c_{i,j}^I}{\sigma_{\hat{w},i}^2}\right) \right], \quad (13)$$

where $c_{i,j}^I$ is the imaginary counterpart of $c_{i,j}^R$.

IV. OPTIMIZATION TECHNIQUE

The optimization problem in equation (4) is a constrained one. We have adopted gradient projection method [15] to solve it. In this optimization formulation, the variables of interests i.e $\alpha_1, \alpha_2 \dots \alpha_N$ are all positive and real-valued quantities. The $Q(x)$ -function is approximated as $Q(x) < \frac{1}{2} \exp(-\frac{x^2}{2})$ [14]. The initial values of α are obtained from the traditional optimized sum-rate CF [3], [5], [16]. It is also noted that the proposed CF is not convex and suffers from multiple minima.

Hence, it is not optimal in our proposed case, but traditional MBER solutions have been shown to outperform others even with steepest descent method [12]. The optimization is iterative and it goes on till an exit condition is reached. There are two exit conditions. The first one is the point till the difference between the previous iteration's CF value and the current one has reached a minimal level defined as ϵ_{CF} . The second one is if it reaches the maximum number of allowed iterations defined as I_{max} . For each iteration, the step size is taken as μ . The update gradient i.e ∇f_i at the i^{th} iteration for α is calculated as follows [17]

$$\nabla f_i \triangleq [\mathbf{A}_{i-1} \nabla P_{e,i-1}(\alpha)] - \mathbf{G}_{i-1} [\mathbf{G}_{i-1}^T \mathbf{G}_{i-1}]^{-1} \mathbf{G}_{i-1} g_{i-1}, \quad (14)$$

where $\mathbf{A}, \mathbf{G}, g$ matrices at the i^{th} iteration are defined as $g = \sum_{j=1}^N [\alpha_j - 1 + R_j - R^{req}]$, $\mathbf{G}_i = \nabla g_i$, $\mathbf{A}_i = \mathbf{I}_N - \mathbf{G}_{i-1} [\mathbf{G}_{i-1}^T \mathbf{G}_{i-1}]^{-1} \mathbf{G}_{i-1}$.

The overall algorithm is described in Algorithm 1.

Algorithm 1 Algorithm to find α^{opt}

- 1: **Initial:** At the Tx, N , R_i^{req}, h_i, σ_i for $i = 1, 2 \dots N$ are given.
 - 2: Obtain initial value of α from sum-rate CF optimization.
 - 3: $k=1$;
 - 4: **while** ($k > 0$)
 - 5: Numerically calculate $\nabla P_{e,i}(\alpha_k)$ and ∇f_i .
 - 6: $\alpha_{k+1} = \alpha_k + \mu \nabla f_i$.
 - 7: Update $\mathbf{A}_i, \mathbf{G}_i, g$.
 - 8: **if** ($|P_e(\alpha_{k+1}) - P_e(\alpha_k)| < \epsilon_{CF}$) or $k > I_{max}$
 - 9: **EXIT while loop**
 - 10: **else** $k = k + 1$; continue
 - 11: **end while**
 - 12: **return** α^{opt} .
-

V. NUMERICAL RESULTS

We study the BER performance of the proposed power NOMA system, where the user power coefficients are derived based on the MBER criterion. We have taken $N = 8$ users with a single BS. We have assumed the channel variances for eight users as $0, -2, -4, -5, -6, -7, -8, -9$ dB, respectively. Let us assume that each user has equal noise variance i.e σ_w^2 , for simplicity purpose. It is also assumed that each user can do a perfect successive-interference-cancellation (SIC) decoding for users present at the lower rank of the i^{th} user. We have considered an uncoded system. We have also considered a minimum rate for each user as 0.2 bits/s/Hz. The received signal-to-noise ratio (SNR) at the i^{th} user is defined as

$$SNR_i \triangleq \frac{\sigma_{h,i}^2 \sigma_x^2 \alpha_i P_T}{\sigma_w^2}. \quad (15)$$

As evident, the channel is assumed to be frequency flat fading. We have chosen $\mu = 10^{-5}$ and $I_{max} = 400$, $\epsilon_{CF} = 10^{-5}$.

With those above configurations, we have plotted the BER performance of our proposed technique for user-1 along with

$$P_c^{i,R}(\alpha) = \Pr \left[\Re(x_i + \frac{w_i}{h_i \sqrt{\alpha_i P_T}} + \frac{1}{\sqrt{\alpha_i}} \sqrt{\alpha_{i-1}}^T \mathbf{x}^{i-1}) \Re(x_i) > 0 \right],$$

$$= \frac{1}{L_i} \sum_{j=1}^{L_i} \Pr \left[\Re(x_i + \frac{w_i}{h_i \sqrt{\alpha_i P_T}} + \frac{1}{\sqrt{\alpha_i}} \sqrt{\alpha_{i-1}}^T \mathbf{x}^{i-1}(j)) \Re(x_i) > 0 | \mathbf{x}^i(j) \right], \quad (9)$$

$$P_c^{i,I}(\alpha) = \frac{1}{L_i} \sum_{j=1}^{L_i} \Pr \left[\Im(x_i + \frac{w_i}{h_i \sqrt{\alpha_i P_T}} + \frac{1}{\sqrt{\alpha_i}} \sqrt{\alpha_{i-1}}^T \mathbf{x}^{i-1}(j)) \Im(x_i) > 0 | \mathbf{x}^i(j) \right], \quad (10)$$

where L_i is the total number of realizations possible with \mathbf{x}^i and $\mathbf{x}^i(j)$ is the j^{th} realization for $j = 1, 2 \dots L_i$.

the conventional sum-rate based CF. For user-1, in Fig. 2, it is observed that at SNR = 16dB, the BER obtained from our proposed algorithm is 3×10^{-3} , while the same BER is obtained at SNR = 20dB for the reference case. Hence, our proposed method gains 4 dB SNR improvement compared to the conventional method. We have obtained similar results for other users. The total rate (sum of all the users) is plotted for our proposed method along with the sum-rate based CF in Fig. 3. It is, however, observed that the spectral efficiency is degraded for the MBER based solution, which is expected. But, the degradation is very small for this configuration of the simulation environment. However, we observe that at lower SNR, the degradation is wider. The key reason is at lower SNR, the noise is making the optimization problem worse. At better optimization can be thought of, though it is out of the scope in this work.

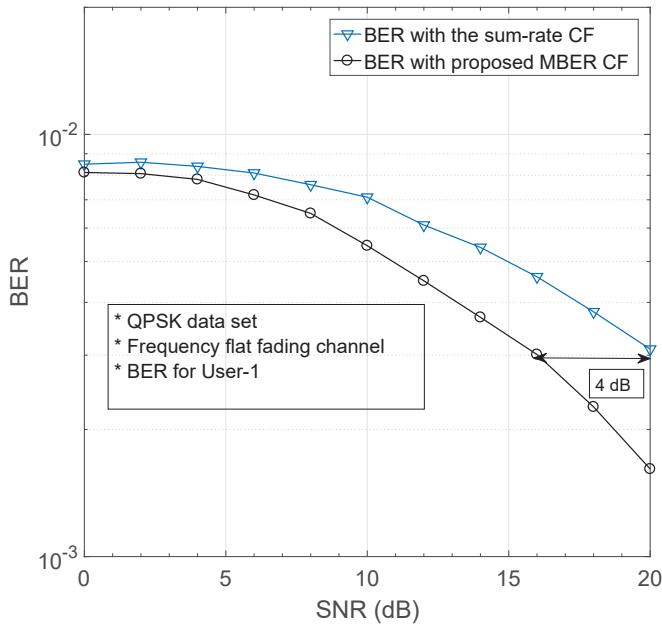


Fig. 2. BER performance for user-1 with the proposed scheme and its comparison with the sum-rate based CF.

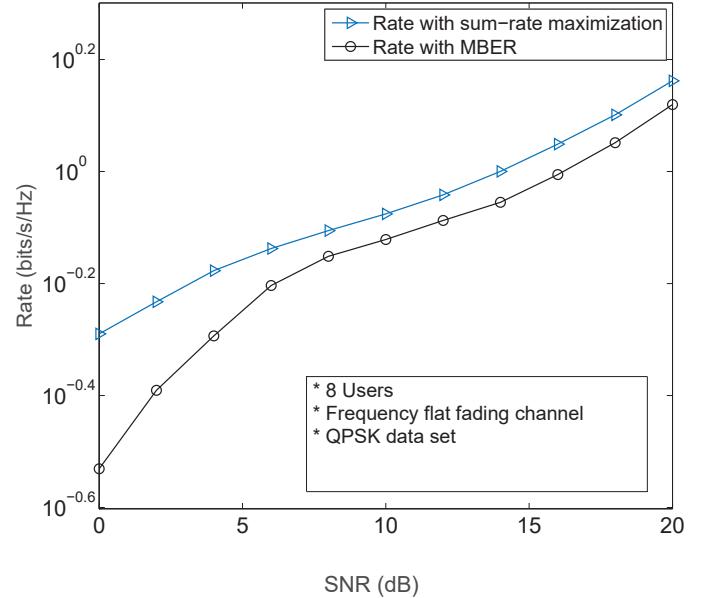


Fig. 3. Total rate performance of the proposed MBER based power NOMA and sum-rate based method.

VI. CONCLUSION

In this treatise, we have proposed the design of power NOMA, where the power coefficients for various users are determined based on the average probability of error for all the users. We have shown the BER improvement observed for various users. This BER performance improvement has been shown experimentally using the proposed solution. The key bottleneck of this valuable work is the computational complexity, which will guide and motivate further development towards this direction.

REFERENCES

- [1] P. Wang, J. Xiao, and L. P, "Comparison of orthogonal and non-orthogonal approaches to future wireless cellular systems," *IEEE Vehicular Technology Magazine*, vol. 1, pp. 4–11, Sept 2006.
- [2] X. Yan, J. Ge, and Y. Zhang, "Researches on non-orthogonal multiple access in multiple-antenna 5g relaying networks," in *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, Oct 2017.

- [3] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE Signal Processing Letters*, vol. 21, pp. 1501–1505, Dec 2014.
- [4] M. Kaneko, H. Yamaura, Y. Kajita, K. Hayashi, and H. Sakai, "Fairness-aware non-orthogonal multi-user access with discrete hierarchical modulation for 5g cellular relay networks," *IEEE Access*, vol. 3, pp. 2922–2938, 2015.
- [5] Y. Wu, L. Qian, H. Mao, W. Lu, H. Zhou, and C. Yu, "Joint channel bandwidth and power allocations for downlink non-orthogonal multiple access systems," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pp. 1–5, Sept 2017.
- [6] J. Wang, H. Xu, L. Fan, B. Zhu, and A. Zhou, "Energy-efficient joint power and bandwidth allocation for noma systems," *IEEE Communications Letters*, vol. 22, pp. 780–783, April 2018.
- [7] M. Kaneko, H. Yamaura, Y. Kajita, K. Hayashi, and H. Sakai, "Fairness-aware non-orthogonal multi-user access with discrete hierarchical modulation for 5g cellular relay networks," *IEEE Access*, vol. 3, pp. 2922–2938, 2015.
- [8] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges," *IEEE Communications Surveys Tutorials*, vol. 19, pp. 721–742, Secondquarter 2017.
- [9] C.-C. Yeh and J. R. Barry, "Adaptive minimum bit-error rate equalization for binary signaling," *IEEE Transactions on Communications*, vol. 48, pp. 1226–1235, July 2000.
- [10] S. Tan, J. Wang, S. X. Ng, S. Chen, and L. Hanzo, "Three-stage turbo mber multiuser beamforming receiver using irregular convolutional codes," *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 1657–1663, May 2008.
- [11] A. Dutta, K. Hari, and L. Hanzo, "Linear Transceiver Design for an Amplify-and-Forward Relay Based on the MBER Criterion," *Communications, IEEE Transactions on*, vol. 62, pp. 3765–3777, Nov 2014.
- [12] A. K. Dutta, K. V. S. Hari, C. R. Murthy, N. B. Mehta, and L. Hanzo, "Minimum error probability mimo-aided relaying: Multihop, parallel, and cognitive designs," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 5435–5440, June 2017.
- [13] S. Vanka, S. Srinivasa, Z. Gong, P. Vizi, K. Stamatou, and M. Haenggi, "Superposition coding strategies: Design and experimental evaluation," *IEEE Transactions on Wireless Communications*, vol. 11, pp. 2628–2639, July 2012.
- [14] J. Proakis, *Digital communications*. McGraw-Hill, 4th ed, 2000.
- [15] J. Rosen., "The gradient projection method for nonlinear programming, Part 2: Nonlinear constraints," *SIAM J. Appl. Math.*, vol. 9, no. 1, pp. 514–553, 1961.
- [16] Z. Sheng, X. Su, and X. Zhang, "A novel power allocation method for non-orthogonal multiple access in cellular uplink network," in *2017 International Conference on Intelligent Environments (IE)*, pp. 157–159, Aug 2017.
- [17] D. H. Luenberger, *Linear and Nonlinear Programming*. Prentice Hall, 1984.

Full-duplex Multi-user Pair Scheduling with Time-selective Fading and Imperfect CSI

Prasanna Raut and Prabhat Kumar Sharma

*Department of Electronics and Communication Engineering
Visvesvaraya National Institute of Technology Nagpur - 440010, India
Email : {rautprasannad@gmail.com, prabhatsharma@ece.vnit.ac.in}*

Abstract—A multi-user full-duplex (FD) two-way communication system with decode-and-forward (DF) relaying protocol is investigated over time-selective fading channels. The effect of imperfect channel state information (CSI) is considered. The fading channel based approach is used to characterize the residual self-interference (RSI) at FD nodes. The outage performance of the considered system is investigated for different scheduling schemes based on the availability of CSI at the relay node. We derive the closed-form tight approximate expressions for the system outage probability assuming independent and non-identically distributed Rayleigh fading channels. Further, the tightness of the approximation presented is verified through Monte-Carlo simulations. Our analysis reveals the significant insights about the impact of time-selective fading, imperfect CSI, and RSI on the performance of the considered system.

Index Terms—decode-and-forward, full-duplex, multi-user, outage probability, scheduling, time-selective fading.

I. INTRODUCTION

The inherent mobility in battery-operated wireless communication devices results in the time-selective channel fading. For physical layer analysis of such systems, various mobility models [1]–[3] like Jakes model [1], Krauss model, UDell model, and Gibb's model [2] were proposed in the literature. The earlier investigations [1], [3], [4] on the effect of time selective fading concluded that the mobility of the communicating wireless nodes severely affects the system performance. Till date, a lot of work have been done considering the time-selective fading channels. Authors in [1] used the first autoregressive model (AR1 model) to represent the outdated channel estimates and studied its effects on the performance of relay selection in amplify-and-forward (AF) cooperative networks. In [3], the effect of relay's mobility was considered with fixed amplification gain and time-selective fading links, and based on pilot signal transmission at the source, a channel estimation method was derived. In [4], authors assumed the perfect estimation at the relay node and at the destination, and studied the effect of time-selective fading on the error rate performance of a single-relay cooperative network.

In advanced wireless communication systems (WCS), the resource scarcity is a common issue, which motivated the sharing of resources among the communicating users. Specifically in multi-user (MU) scenario, the relay sharing has been

This research was supported by the Early Career Research grant (*ECR/2016/000196*) funded by the Science and Engineering Research Board, Government of India.

emerged as a promising solution. In such systems, relay selects, based on some well defined mechanisms, a pair of users which communicate with each other. This process is known as the user scheduling. The multi-user scheduling systems have been widely explored in the recent literature [5]–[8]. As demonstrated in [6], [8], compared to the single source single destination (1 – 1¹) decode-and-forward (DF) relay system, full-duplex (FD) $N - N$ multi-user relay system can give the advantage of multi-user diversity gain by adopting different scheduling schemes. The use of FD relays however imposes a critical concern in the form of residual self-interference (RSI). The RSI has been characterized as random fading [9] and as constant power [6] in the recent literatures. For the purpose of selecting the user pair, the authors in [6] investigated different scheduling schemes such as random scheduling, max-min scheduling, optimal scheduling for FD MU two-way relaying (TWR) system assuming perfect estimation at the relay and the destination. However, aforementioned scheduling schemes in [6]–[8] are based on the best channel criteria to maximize the system throughput. Thus later, equivalent fair scheduling schemes which use modified “relative channel strength” criteria were proposed in [8].

To the best of authors' knowledge, none of the recent works mentioned earlier related to FD TWR focuses on the performance of the system with time-selective fading. In addition, the analysis of FD MU TWR system with Rician faded RSI is also an open research problem. The characterization of RSI as Rician fading is experimentally proven in [10]. Motivated by this, in this paper, a FD TWR multi-user system with physical-layer network coding based DF relaying protocol is considered. All the nodes are assumed to be mobile. All the channels but the self-interference channels are assumed to be time-selective independent and non-identically distributed (i.n.i.d) complex Gaussian (Rayleigh envelope) random variable (RV) which has been widely used for characterizing multi-path components with no line-of-sight in WCS. The instantaneous RSI at the relay and the user nodes is characterized as Rician distributed RV.

The remainder of this paper is organized as follows. After introducing the system model of the proposed multi-user FD-TWR system in Section II, the closed-form expressions for the outage probability of the considered scheduling schemes are

¹ $N - N$ multi-user model refers to N source and N destinations.

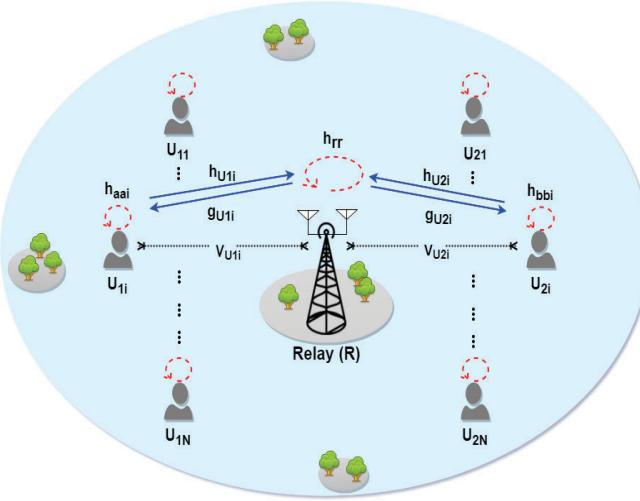


Fig. 1: The $N - N$ FD multi-user TWR system.

derived in section III. Numerical simulations are presented in section IV, and section V concludes this paper. The notations used in this paper are as follows. The index $\{i\} \in \{1, 2, \dots, N\}$, $\{pq\} \in \{\{U_{1i}\}, \{U_{2i}\}\}$, $\{xy\} \in \{\{aa_i\}, \{bb_i\}, \{rr\}\}$. $Pr\{\cdot\}$ and $\mathbb{E}\{\cdot\}$ denote the probability and the statistical expectation operation, respectively. $|\cdot|$ denotes the absolute value of a complex scalar. Further, $f(\cdot)$, $\mathcal{F}(\cdot)$ denote the probability density function (PDF) and the cumulative distribution function (CDF), respectively.

II. SYSTEM MODEL

We consider the $N - N$ multi-user system model with $2N$ mobile FD users as depicted in Fig. 1, where the i th selected user pair (U_{1i}, U_{2i}) post scheduling desires to exchange information via a bidirectional FD relay node R . All nodes are assumed to work as mobile terminals and operate in FD mode. We assume that due to the large separation and strong shadowing effects, the direct link between U_{1i} and U_{2i} does not exist. Further, in each time slot, the FD relay node R can only accommodate one pair of users to exchange data, and in every time slot only selected pair is active while all the other user pairs are in idle mode till the next time slot.

A. Channel Model

The channel coefficient between the node U_{1i} and R (R and U_{1i}) is represented as $h_{U_{1i}}$ ($g_{U_{1i}}$) and that between the node U_{2i} and R (R and U_{2i}) is represented as $h_{U_{2i}}$ ($g_{U_{2i}}$). Further, the relative speed between U_{1i} and relay node is denoted by $V_{U_{1i}}$ and that between U_{2i} and relay node is denoted by $V_{U_{2i}}$. Because of the user nodes' mobility, all the block fading channels $h_{U_{1i}}$, $g_{U_{1i}}$, $h_{U_{2i}}$, and $g_{U_{2i}}$ are time selective and are subject to independent and non-identically distributed (i.n.i.d) complex Gaussian fading coefficients (Rayleigh envelope) with parameter $\omega_{pq} = \mathbb{E}\{|h_{pq}|^2\}$ and $\bar{\omega}_{pq} = \mathbb{E}\{|g_{pq}|^2\}$. For the Rayleigh distributed channel t_{pq} , $|t_{pq}|^2$ is exponentially distributed with mean $\lambda_{pq} = 1/\omega_{pq}$ for channel h_{pq} and with mean $\tau_{pq} = 1/\bar{\Omega}_{pq}$ for channel g_{pq} with $\{t\} \in \{h, g\}$.

Thus following Jake's model, the time adjacent channel gains of fading links are considered uncorrelated with correlation parameter ρ_{pq} which is given by $\rho_{pq} = \mathcal{J}_0\left(\frac{2\pi f_c V_{pq}}{R_s c}\right)$. The parameter f_c is carrier frequency, c is speed of light, R_s is transmission symbol rate and \mathcal{J}_0 represents zeroth-order Bessel function of first kind [11]. The time-selective fading links h_{pq} , g_{pq} are modelled by the first order autoregressive (AR1) process as [1]

$$h_{pq}(\tau_1) = \rho_{pq}h_{pq}(\tau_2) + e_{pq}(\tau_1)\sqrt{1 - \rho_{pq}^2}, \\ g_{pq}(\tau_1) = \rho_{pq}g_{pq}(\tau_2) + \tilde{e}_{pq}(\tau_1)\sqrt{1 - \rho_{pq}^2}, \quad (1)$$

where $t_{pq}(\tau_1)$, $t_{pq}(\tau_2)$, $\{t\} \in \{h, g\}$ are the channel gains over two adjacent time positions τ_1 and τ_2 and are distributed as zero mean circularly symmetric complex Gaussian (ZM-CSCG). Said differently, $h_{pq}(\tau) \sim \mathcal{CN}(0, \sigma_{h_{pq}}^2)$, and $g_{pq}(\tau) \sim \mathcal{CN}(0, \sigma_{g_{pq}}^2)$. The random processes $e_{pq}(\tau_1)$, $\tilde{e}_{pq}(\tau_1)$ are the varying-component of the associated link distributed as $e_{pq}(\tau_1) \sim \mathcal{CN}(0, \sigma_{e_{pq}}^2)$, and $\tilde{e}_{pq}(\tau_1) \sim \mathcal{CN}(0, \sigma_{\tilde{e}_{pq}}^2)$, respectively. In this paper, we assume k th signalling position and is repeatedly used for further analysis in the rest of paper.

Further h_{aa_i} , h_{bb_i} , h_{rr} represents the RSI channels at the user nodes U_{1i} , U_{2i} and the FD relay node R , respectively. Since, the RSI channels exist between the transmit and receive antenna of same node, the relative velocity between the antennas will be zero. Hence, RSI channels are not assumed to be time-selective. According to the results of experimental characterization in [10], the RSI channel should be modeled with Rician distribution with a lower value of K-factor. Thus, in this paper the RSI channels are assumed to be i.n.i.d. Rician distributed with parameter $\Omega_{xy} + \vartheta_{xy}^2 = \mathbb{E}\{|h_{xy}|^2\}$ and $K_{xy} = \vartheta_{xy}^2/\Omega_{xy}$ with $\{xy\} \in \{\{aa_i\}, \{bb_i\}, \{rr\}\}$ where the scale parameter Ω_{xy} is defined as the total power received in all paths while the shape parameter K_{xy} is defined as the ratio of the power contributions by line-of-sight path to the remaining multi paths. Based on the availability of CSI at the relay node, the random scheduling, absolute SINR-based scheduling, normalized SINR-based scheduling scheme are adopted to select the user pair.

B. Transmission Protocol

In this paper, the physical-layer network coding (PNC) based [12] TW DF protocol is adopted for the processing of signals. First, from selected user pair i , the users U_{1i} and U_{2i} simultaneously send signals $X_{U_{1i}}(k)$ and $X_{U_{2i}}(k)$ to the relay node R , respectively. We denote the transmission power of user nodes U_{1i} and U_{2i} as $P_{U_{1i}} = \mathbb{E}\{|X_{U_{1i}}|^2\}$ and $P_{U_{2i}} = \mathbb{E}\{|X_{U_{2i}}|^2\}$, respectively. As the considered fading links are time-selective, we assume that the relay is not able to track the channel coefficients instantaneously. Hence relay estimates the channel coefficients over first signalling period of each transmitted block as $\hat{h}_{pq}(1)$ and $\hat{g}_{pq}(1)$ using a pilot signal

as discussed in [1], [3]. Thus, the expression of h_{pq} , in (1) modifies as in [1], to

$$h_{pq}(k) = \rho_{pq}^{k-1} \hat{h}_{pq}(1) + \underbrace{\sqrt{1 - \rho_{pq}^2} \sum_{m=1}^{k-1} \rho_{pq}^{k-m-1} e_{pq}(m)}_{=\phi_{pq}(k)} + \rho_{pq}^{k-1} h_{\epsilon_{pq}}(1), \quad (2)$$

where \hat{h}_{pq} , \hat{g}_{pq} are the estimated channel coefficients distributed as ZM-CSCG with mean zero and variances $\sigma_{\hat{h}_{pq}}^2$, $\sigma_{\hat{g}_{pq}}^2$, respectively. Further $h_{\epsilon_{pq}}$, $g_{\epsilon_{pq}}$ are the estimation errors and are assumed to be ZM-CSCG distributed random variables with variances $\sigma_{h_{\epsilon_{pq}}}^2$, $\sigma_{g_{\epsilon_{pq}}}^2$, respectively. Using (2), the received signal at the relay node in first time slot is given by

$$\begin{aligned} Y_R(k) &= \underbrace{\rho_{U_{1i}}^{k-1} \hat{h}_{U_{1i}}(1) X_{U_{1i}}(k)}_{\text{desired signal from } U_{1i}} + \underbrace{\rho_{U_{2i}}^{k-1} \hat{h}_{U_{2i}}(1) X_{U_{2i}}(k)}_{\text{desired signal from } U_{2i}} \\ &\quad + \underbrace{\rho_{U_{1i}}^{k-1} h_{\epsilon_{U_{1i}}}(1) X_{U_{1i}}(k)}_{\text{imperfect CSI noise}} + \underbrace{\rho_{U_{2i}}^{k-1} h_{\epsilon_{U_{2i}}}(1) X_{U_{2i}}(k)}_{\text{imperfect CSI noise}} \\ &\quad + \underbrace{\phi_{U_{1i}}(k) X_{U_{1i}}(k)}_{\text{nodes' mobility noise}} + \underbrace{\phi_{U_{2i}}(k) X_{U_{2i}}(k)}_{\text{nodes' mobility noise}} \\ &\quad + \underbrace{n_R(k)}_{\text{AWGN noise}} + \underbrace{h_{rr}(k) \hat{X}_R(k)}_{\text{self interference noise}}, \end{aligned} \quad (3)$$

where \hat{X} is the decoded version of signal received in the previous time instant; $n_R(k)$ is the additive white Gaussian noise (AWGN) at R with $n_R(k) \sim \mathcal{N}(0, N_0)$. The effective noise at relay node consist of all the terms in (3) except the first two. Since $h_{\epsilon_{pq}}$, $e_{pq}(m)$, and $n_R(k)$ are ZM-CSCG processes, the sum of second, third and fourth noise terms in (3) is also ZM-CSCG with effective noise power

$$\begin{aligned} \sigma_{R_{eff}}^2 &= (1 - \rho_{U_{1i}}^{2(k-1)}) \sigma_{e_{U_{1i}}}^2 P_{U_{1i}} + (1 - \rho_{U_{2i}}^{2(k-1)}) \sigma_{e_{U_{2i}}}^2 P_{U_{2i}} \\ &\quad + \rho_{U_{1i}}^{2(k-1)} \sigma_{h_{\epsilon_{U_{1i}}}}^2 P_{U_{1i}} + \rho_{U_{2i}}^{2(k-1)} \sigma_{h_{\epsilon_{U_{2i}}}}^2 P_{U_{2i}} + N_0. \end{aligned}$$

The relay node R jointly decodes the received signals $X_{U_{1i}}(k)$ and $X_{U_{2i}}(k)$. Then it broadcasts the re-encoded signal $\hat{X}_R(k) = X_{U_{1i}}(k) \oplus X_{U_{2i}}(k)$ to nodes U_{1i} and U_{2i} after performing bit-level PNC [12] with transmission power $P_R = \mathbb{E}\{|\hat{X}_R|^2\}$. Here, $A \oplus B$ represents the bit-level exclusive (XOR) operation. Similar to (2), expanding g_{pq} for k th signaling position, received signals $Y_{U_{1i}}$ and $Y_{U_{2i}}$ at nodes U_{1i} and U_{2i} , respectively, can be written as

$$\begin{aligned} Y_{U_{1i}}(k) &= \rho_{U_{1i}}^{k-1} \hat{g}_{U_{1i}}(1) \hat{X}_R(k) + \rho_{U_{1i}}^{k-1} g_{\epsilon_{U_{1i}}}(1) \hat{X}_R(k) \\ &\quad + \tilde{\phi}_{U_{1i}}(k) \hat{X}_R(k) + n_{U_{1i}}(k) + h_{aa_i}(k) X_{U_{1i}}(k), \\ Y_{U_{2i}}(k) &= \rho_{U_{2i}}^{k-1} \hat{g}_{U_{2i}}(1) \hat{X}_R(k) + \rho_{U_{2i}}^{k-1} g_{\epsilon_{U_{2i}}}(1) \hat{X}_R(k) \\ &\quad + \tilde{\phi}_{U_{2i}}(k) \hat{X}_R(k) + n_{U_{2i}}(k) + h_{bb_i}(k) X_{U_{2i}}(k) \end{aligned} \quad (4)$$

with effective noise variances at node U_{1i} and U_{2i} given by

$$\begin{aligned} \sigma_{U_{1i}eff}^2 &= (1 - \rho_{U_{1i}}^{2(k-1)}) \sigma_{e_{U_{1i}}}^2 P_R + \rho_{U_{1i}}^{2(k-1)} \sigma_{g_{\epsilon_{U_{1i}}}}^2 P_R + N_0, \\ \sigma_{U_{2i}eff}^2 &= (1 - \rho_{U_{2i}}^{2(k-1)}) \sigma_{e_{U_{2i}}}^2 P_R + \rho_{U_{2i}}^{2(k-1)} \sigma_{g_{\epsilon_{U_{2i}}}}^2 P_R + N_0, \end{aligned}$$

where $n_{U_{1i}}(k)$ and $n_{U_{2i}}(k)$ represent AWGN at nodes U_{1i} and U_{2i} , respectively with mean zero and variance N_0 . It should be noted that, if the relative speed between i th selected user and R is zero then $\rho_{pq}=1$ i.e., the case when nodes are stationary. Further, for perfect estimation, the value of $\sigma_{h_{\epsilon_{pq}}}^2$ is zero.

With the help of physical-layer signal model, the respective signal-to-interference-plus-noise-ratios (SINRs) for the forward links and the backward links using (3) and (4) can be written as

$$\begin{aligned} \gamma_{U_{1i}R} &= \frac{\rho_{U_{1i}}^{2(k-1)} |\hat{h}_{U_{1i}}(1)|^2 P_{U_{1i}}}{|h_{rr}|^2 P_R + \sigma_{R_{eff}}^2}, \quad \gamma_{RU_{2i}} = \frac{\rho_{U_{2i}}^{2(k-1)} |\hat{g}_{U_{2i}}(1)|^2 P_R}{|h_{bb_i}|^2 P_{U_{2i}} + \sigma_{U_{2i}eff}^2}, \\ \gamma_{RU_{1i}} &= \frac{\rho_{U_{1i}}^{2(k-1)} |\hat{g}_{U_{1i}}(1)|^2 P_R}{|h_{aa_i}|^2 P_{U_{1i}} + \sigma_{U_{1i}eff}^2}, \quad \gamma_{U_{2i}R} = \frac{\rho_{U_{2i}}^{2(k-1)} |\hat{h}_{U_{2i}}(1)|^2 P_{U_{2i}}}{|h_{rr}|^2 P_R + \sigma_{R_{eff}}^2}, \end{aligned} \quad (5)$$

where SINRs at selected nodes U_{1i} and U_{2i} are represented as $\gamma_{RU_{1i}}$ and $\gamma_{RU_{2i}}$, respectively; $\gamma_{U_{1i}R}$ and $\gamma_{U_{2i}R}$ are the SINRs of link from node U_{1i} to relay R and node U_{2i} to relay R , respectively. Finally, the sum SINR [5], [6], [13] at relay R is represented as $\gamma_{sum} = \gamma_{U_{1i}R} + \gamma_{U_{2i}R}$.

III. OUTAGE ANALYSIS

The outage event is said to be occurred if the i th selected link has achievable rate (C_i) less than that of minimum required rate (r_i) [13]. i.e., $\mathcal{P}_i \triangleq \Pr\{C_i < r_i \mid C_i\}$, where $C_i = \log_2(1 + \gamma_i)$. Furthermore, we assume the asymmetric traffic requirements so that the forward and the backward links have non-identical minimum threshold rates r_1 and r_2 , respectively. With the help of Shannon's capacity formula, the overall outage event for the selected user pair is expressed as [5], [6], [13]

$$\begin{cases} \min\{\log_2(1 + \gamma_{U_{1i}R}), \log_2(1 + \gamma_{RU_{2i}})\} < r_1, & \text{or} \\ \min\{\log_2(1 + \gamma_{U_{2i}R}), \log_2(1 + \gamma_{RU_{1i}})\} < r_2, & \text{or} \\ \log_2(1 + \gamma_{U_{1i}R} + \gamma_{U_{2i}R}) < r_1 + r_2. \end{cases} \quad (6)$$

Using $R_1 = 2^{r_1} - 1$, $R_2 = 2^{r_2} - 1$ and $R = 2^r - 1$ with $r = r_1 + r_2$, we can redefine the outage event in (6) as

$$\min\{\gamma_{\bar{a}r}, \gamma_{\bar{b}r}\} < 1 \quad \text{or} \quad \min\{\gamma_{\bar{r}a}, \gamma_{\bar{r}b}\} < 1 \quad \text{or} \quad \gamma_{\bar{s}um} < 1 \quad (7)$$

where $\gamma_{\bar{a}r} = \frac{\gamma_{U_{1i}R}}{R_1}$, $\gamma_{\bar{b}r} = \frac{\gamma_{RU_{2i}}}{R_1}$, $\gamma_{\bar{r}a} = \frac{\gamma_{RU_{1i}}}{R_2}$, $\gamma_{\bar{r}b} = \frac{\gamma_{U_{2i}R}}{R_2}$, and $\frac{\gamma_{U_{1i}R} + \gamma_{U_{2i}R}}{R} = \frac{\gamma_{sum}}{R} = \gamma_{\bar{s}um}$. From (7), the end-to-end SINR Φ_i of the i th selected user pair is

$$\Phi_i = \min\{\gamma_{\bar{a}r}, \gamma_{\bar{b}r}\} \quad \text{or} \quad \min\{\gamma_{\bar{r}a}, \gamma_{\bar{r}b}\} \quad \text{or} \quad \gamma_{\bar{s}um}. \quad (8)$$

In (7), let M denotes $\min\{\gamma_{\bar{a}r}, \gamma_{\bar{b}r}\}$, N denotes $\min\{\gamma_{\bar{r}a}, \gamma_{\bar{r}b}\}$ and O denotes $\gamma_{\bar{s}um}$, then the exact outage probability of the i th scheduled pair P_{out}^i is expressed as

$$\begin{aligned} P_{out}^i &= \Pr\{\Phi_i < 1\}, \\ P_{out}^i &= \Pr\{(M, N, O) : M < 1 \cup N < 1 \cup O < 1\}. \end{aligned} \quad (9)$$

From (5) and (7), it is apparent that the events M , N , and O in (9) are dependent events and with Rician distributed RSI (h_{aa_i} , h_{bb_i} and h_{rr}) it is difficult to tackle the mathematical

analysis of (9). Hence we approximate the outage event in (9) and derive the approximated closed form expression of outage probability. In (9), it is observed through Monte Carlo simulations that, event O does not affect P_{out}^i severely in the low SINR region. Hence we neglect event O and the lower bound of outage probability P_{out}^{iL} is obtained from (9) as done in [14], as

$$P_{out}^i \geq P_{out}^{iL} = Pr\{\Phi_i^L < 1\} = Pr\{(M, N) : M < 1 \cup N < 1\} \quad (10)$$

where $\Phi_i^L = \min\{\gamma_{\bar{a}r}, \gamma_{\bar{b}r}\} \cup \min\{\gamma_{\bar{r}a}, \gamma_{\bar{b}r}\}$ with M and N are dependent events conditioned on $|h_{rr}|^2$. The tightness of the approximation is verified through the exact Monte-Carlo simulations and is shown in numerical results section. Further simplifying (10) we obtain

$$\begin{aligned} P_{out}^i &\geq P_{out}^{iL} = Pr\{(M, N) : M < 1\} + Pr\{(M, N) : N < 1\} \\ &\quad - Pr\{(M, N) : M < 1 \cap N < 1\} \\ P_{out}^{iL} &= \mathcal{F}_{\Phi_i^L}(1) = \mathcal{F}_M(1) + \mathcal{F}_N(1) - \mathcal{F}_{M \cap N}(1), \end{aligned} \quad (11)$$

where $\mathcal{F}_{M \cap N}(1)$, $\mathcal{F}_M(1)$, and $\mathcal{F}_N(1)$ are given by (16), (18), and (19), respectively.

Proof: See Appendix A. ■

A. Multi-user Scheduling

In this subsection, we discuss different multi-user scheduling schemes and their criteria for selecting i th user.

1) *Random Scheduling:* For the blind case, where the instantaneous CSI is not available at relay node, we adopt random (RND) scheduling scheme where relay nodes randomly selects a user pair among N pairs. Since in every time slot, scheduling of all user pairs is equally probable, by the total probability theorem, the outage probability of system with the RND scheduling for i.n.i.d. channels can be written as

$$P_{out}^{RND} \approx \frac{1}{N} \sum_{i=1}^N P_{out}^{iL} \quad (12)$$

where P_{out}^{iL} is given by (11).

2) *Absolute SINR-Based Scheduling:* In presence of full CSI of h_{pq} and g_{pq} at relay node, we adopt absolute (ABS) SINR-based scheduling, in which the user pair i with the maximum instantaneous SINR Φ_i^L is chosen in each time slot. Said differently, $i = \arg \max_{i \in N} \{\Phi_i^L\}$. The system outage probability for ABS SINR-based Scheduling is given as [8]

$$P_{out}^{ABS} \approx \max_{i \in N} \left\{ \mathcal{F}_{\Phi_i^L}(1) \right\} \approx \prod_{i=1}^N P_{out}^{iL}. \quad (13)$$

3) *Normalized SINR-Based Scheduling:* The selection of user pair based on maximum instantaneous SINR can lead to a biased scheduling of the system resources among all the users. In system with this scheduling scheme, the users with the best SINR on average will dominate the rest of the users. Thus to achieve fairness among all users, in normalized (NRM) SINR-based scheduling, modified selection criteria of users based on the relative SINR is considered where the user pair

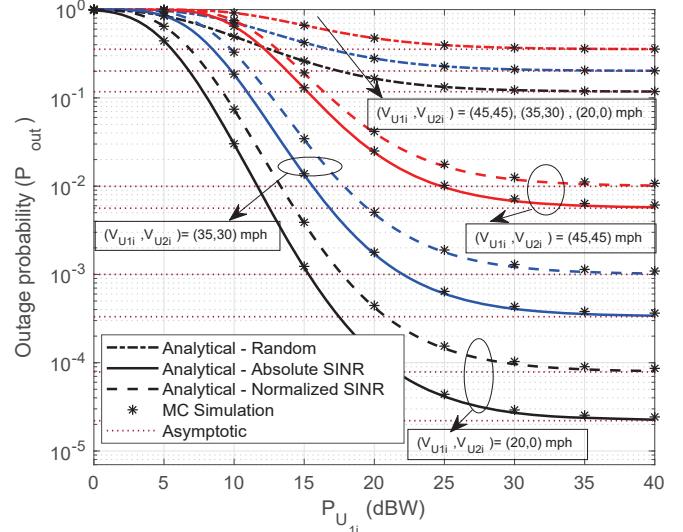


Fig. 2: Outage probability of system versus transmit user node power $P_{U_{1i}}$ with $P_{U_{2i}} = P_R = P_{U_{1i}}$, with nodes' mobility.

i with the maximum normalized SINR $\Phi_i^L/\bar{\Phi}_i^L$ is chosen. i.e., $i = \arg \max_{i \in N} \left\{ \frac{\Phi_i^L}{\bar{\Phi}_i^L} \right\}$ where $\bar{\Phi}_i^L$ is the average value of Φ_i^L for the i th user pair. Using this criteria, the system outage probability for NRM SINR-based Scheduling is given as [8]

$$P_{out}^{NRM} \approx \mathcal{F}_{\Phi_i^L}(1), \quad (14)$$

where $\mathcal{F}_{\Phi_i^L}(\cdot)$ is the CDF of SINR of the i th user pair post normalized SINR-based scheduling and is given by [8]

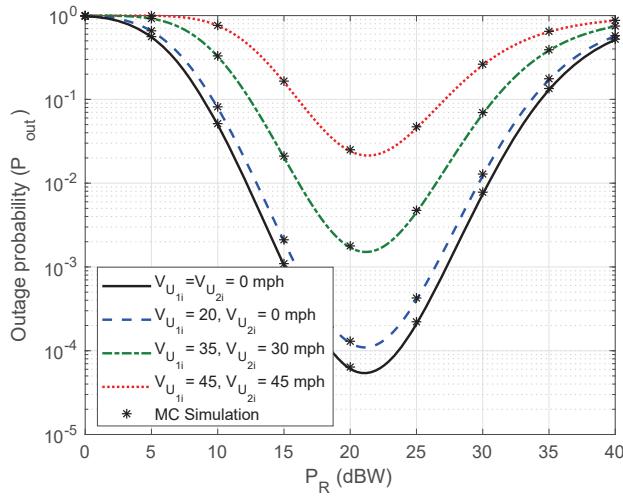
$$\mathcal{F}_{\Phi_i^L}(1) = \frac{1}{N} \sum_{i=1}^N \left(\mathcal{F}_k \left(\frac{\Phi_i^L}{\bar{\Phi}_i^L} \right) \right)^N, \quad (15)$$

with $\{k\} = \{1, 2, \dots, N\}$ and $k \neq i$. The term $\mathcal{F}_k \left(\frac{\Phi_i^L}{\bar{\Phi}_i^L} \right)$ is derived in similar way as done earlier in appendix A. The expression for $\mathcal{F}_k \left(\frac{\Phi_i^L}{\bar{\Phi}_i^L} \right)$ is obtained by replacing $\bar{\gamma}_{\bar{a}r}$, $\bar{\gamma}_{\bar{r}b}$, $\bar{\gamma}_{\bar{r}a}$, and $\bar{\gamma}_{\bar{b}r}$ with $\bar{\Phi}_i^L$ in the expression P_{out}^{iL} given by (12). Further, to obtain the asymptotic floor for the outage probability, we modify (11) with $P_{U_{1i}}, P_{U_{2i}}, P_R \rightarrow \infty$ and the expression is omitted due to space limitation.

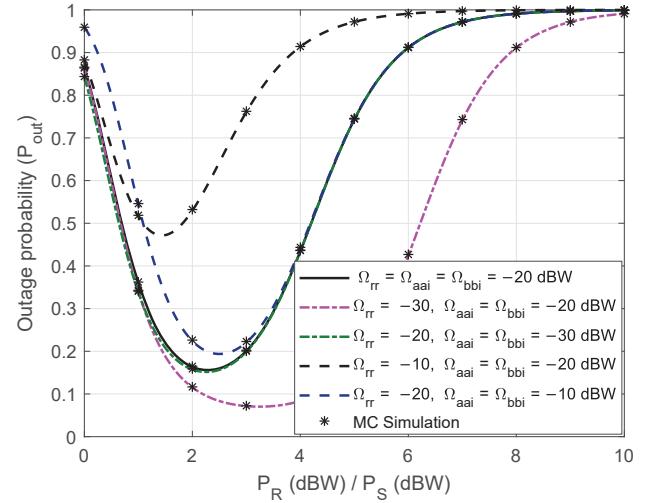
IV. NUMERICAL RESULTS AND DISCUSSIONS

In this section we present the analytical and simulated results. The parameters ω_{pq} and $\bar{\omega}_{pq}$ are generated between 0.5 and 1, randomly. Minimum required rates $r_1 = r_2 = 1$ bps/Hz, $N_0 = 1$ and number of user pairs $N = 5$. Monte Carlo (MC) simulations for the exact outage probability are presented to verify the approximated theoretical analysis and are performed over 10^8 samples.

In Fig. 2, we set $P_R = P_{U_{2i}} = P_{U_{1i}}$, the RSI powers Ω_{xy_i} equal to -20 dBW (Power in decibels with respect to 1W) and Rician factor $K_{xy_i} = 1$ dBW. From Fig. 2, we note that time-selective fading critically affects the system performance and higher the nodes' mobility, poorer is the outage performance.



(a) Impact of the nodes mobility on the outage minima



(b) Impact of the RSI on the outage minima

Fig. 3: P_{out} versus the relay transmit powers in dBW and (a), (b) corresponds to ABS SINR based, RND scheduling, respectively.

In addition we note that the outage probability of system decreases first and then saturates after certain value. Any further increase in the transmit power does not improve the performance which is limited by the asymptotic floors. For different nodes' mobility scenario, we observed that mobile relays ($V_{U_{1i}} = V_{U_{2i}} = 45$ mph) severely affects the performance of the system as compared to mobile user nodes ($V_{U_{1i}} = 35$, $V_{U_{2i}} = 30$ mph). Further, in terms of the performance of the various scheduling schemes, we note that the absolute SINR-based scheduling outperforms the normalized SINR-based scheduling though the later one is more fair.

In Fig. 3a $P_{U_{1i}} = P_{U_{2i}} = P_S$ is set equal to 20 dBW with $\Omega_{xy_i} = -20$ dBW and $K_{xy_i} = 1$ dBW. From this figure, we observe that the outage probability first decreases and then increases with P_R and finally approaches to unity. This corroborates that the high transmit power strengthens the RSI which further reduces the SINR. From the comparison of outage probability, we obtain the insight that the outage minima is independent of the $V_{U_{1i}}$, $V_{U_{2i}}$. In other words, with given P_S , the outage minima occurs at the same P_R irrespective of the nodes mobility. However as can be seen from Fig. 3a, it is also observed that lower mobility results in lower outage probability.

Fig. 3b investigates the outage probability of the system for random scheduling with the ratio of transmit powers of relay node to user nodes for different values of RSI powers Ω_{xy_i} assuming $P_{U_{1i}} = P_{U_{2i}} = P_S$. We set $K_{xy_i} = 1$ dBW and $V_{U_{1i}} = V_{U_{2i}} = 5$ mph. From Fig. 3b we conclude that with change in P_R with respect to P_S by the factor between 0 - 10, the outage probability of system gets its minima for the least RSI condition. However, this minimum point gets shifted for the different RSI conditions at the relay and user nodes unlike the case shown in Fig. 3a. This indicates that the minimum outage performance can be achieved by properly selecting the transmit powers of relay and users. Further, the change in the RSI level of user nodes does not affect the outage probability of system. As shown in Fig. 3b, the outage performance of system for change in K_{aai} , K_{bbi} and Ω_{aai} , Ω_{bbi} of the user nodes is almost similar and overlapping and the outage minima is obtained around $P_R/P_S = 2.5$. In contrast, the change in K_{rr} and Ω_{rr} affect system severely and the point of outage minima varies for different values of K_{rr} and Ω_{rr} . This indicates the dominance of RSI at relay node as compared to that at user nodes in a DF FD relaying system.

$$\begin{aligned} \mathcal{F}_{M \cap N}(1) = & \mathcal{F}_{\gamma_{ra}}(1)\mathcal{F}_{\gamma_{rb}}(1) + (1 - \mathcal{F}_{\gamma_{rb}}(1))\mathcal{F}_{\gamma_{ar}}(1)\mathcal{F}_{\gamma_{ra}}(1) + (1 - \mathcal{F}_{\gamma_{ra}}(1))\mathcal{F}_{\gamma_{rb}}(1)\mathcal{F}_{\gamma_{br}}(1) - (1 - \mathcal{F}_{\gamma_{ra}}(1) - \mathcal{F}_{\gamma_{rb}}(1)) \\ & + \mathcal{F}_{\gamma_{ra}}(1)\mathcal{F}_{\gamma_{rb}}(1) \left(1 - \mathcal{F}_{\gamma_{ar}}(1) - \mathcal{F}_{\gamma_{br}}(1) - \frac{\lambda_{rr}e^{-(\lambda_{rr}v_{rr}^2)}e^{\left(\frac{-c_2\lambda_{rr}}{(1+c_2\lambda_{rr})}\left[\frac{1}{\gamma_{ar}} + \frac{1}{\gamma_{br}}\right]\right)}e^{\left(\frac{\lambda_{rr}v_{rr}^2}{1+\left(\frac{1}{(1+c_2\lambda_{rr})}\left[\frac{1}{\gamma_{ar}} + \frac{1}{\gamma_{br}}\right]\right)}\right)}}{1 + \left(\frac{1}{(1+c_2\lambda_{rr})}\left[\frac{1}{\gamma_{ar}} + \frac{1}{\gamma_{br}}\right]\right)} \right) \end{aligned} \quad (16)$$

V. CONCLUSIONS

This paper studied the outage performance of a FD TWR network system under multi-user environment with nodes mobility. The approximated closed-form outage probability expressions were derived for the absolute and normalized SINR based scheduling schemes with time-selective Rayleigh distributed i.n.i.d. channels. This paper also compared the outage probability of system for three scheduling schemes and it was observed that the time-selective fading channels can critically affect the outage performance of system. Further, it was also concluded that the mobile relay terminal affects system more severely as compared to mobile user nodes.

APPENDIX A PROOF OF EQ. (11)

In (5), let, $c_1 = P_{U_{1i}} \rho_{U_{1i}}^{2(k-1)} / P_R$, $c'_1 = P_{U_{2i}} \rho_{U_{2i}}^{2(k-1)} / P_R$, $c_2 = \sigma_{R_{eff}}^2 / P_R$, $c_3 = P_R \rho_{U_{1i}}^{2(k-1)} / P_{U_{1i}}$, $c'_3 = P_R \rho_{U_{2i}}^{2(k-1)} / P_{U_{2i}}$, $c_4 = \sigma_{U_{1i}eff}^2 / P_{U_{1i}}$, $c'_4 = \sigma_{U_{2i}eff}^2 / P_{U_{2i}}$. The PDF of $|h_{pq}|^2$, $|g_{pq}|^2$, $\{pq\} \in \{U_{1i}, U_{2i}\}$, follows exponential distribution. The PDF of $|h_{xy}|^2$, $\{xy\} \in \{\{aa_i\}, \{bb_i\}, \{rr\}\}$, is given by

$$f_{|h_{xy}|^2}(t) = \frac{1}{\Omega_{xy}} e^{\left(\frac{-(t+\nu_{xy}^2)}{\Omega_{xy}}\right)} I_0\left(\frac{2\nu_{xy}\sqrt{t}}{\Omega_{xy}}\right),$$

for $t \geq 0$ where $I_0(\cdot)$ is the modified Bessel function of the first kind [11]. The CDF of $\gamma_{\bar{ar}}$ is calculated as

$$\begin{aligned} \mathcal{F}_{\gamma_{\bar{ar}}}(1) &= \mathbb{E}_{|h_{rr}|^2} \left[\Pr \left\{ |h_{U_{1i}}|^2 \leq \frac{R_1(|h_{rr}|^2 + c_2)}{c_1} \right\} \right] \\ &= \mathbb{E}_{|h_{rr}|^2} \left[\int_0^{\left(\frac{R_1(|h_{rr}|^2 + c_2)}{c_1} \right)} f_{|h_{U_{1i}}|^2}(t) dt \right] \\ &= \int_0^{\infty} \left(1 - e^{-\left(\frac{\lambda_{U_{1i}} R_1(|h_{rr}|^2 + c_2)}{c_1} \right)} \right) f_{|h_{rr}|^2}(Z) dZ \\ \mathcal{F}_{\gamma_{\bar{ar}}}(1) &= 1 - \frac{e^{-(\lambda_{rr} \nu_{rr}^2)} e^{\left(\frac{-c_2 \lambda_{rr}}{(1+c_2 \lambda_{rr}) \bar{\gamma}_{\bar{ar}}} \right)} e^{\left(\frac{\lambda_{rr} \nu_{rr}^2}{1+(1+c_2 \lambda_{rr}) \bar{\gamma}_{\bar{ar}}} \right)}}}{1 + \frac{1}{(1+c_2 \lambda_{rr}) \bar{\gamma}_{\bar{ar}}}} \quad (17) \end{aligned}$$

where $\bar{\gamma}_{\bar{uv}}$ is the average SINR of the uv link and is defined as $\bar{\gamma}_{\bar{uv}} = \mathbb{E}\{\gamma_{\bar{uv}}\}$ with $\{\bar{uv}\} \in \{\bar{ar}, \bar{rb}, \bar{ra}, \bar{br}\}$. The term $\bar{\gamma}_{\bar{ar}}$ in (17) is given by

$$\bar{\gamma}_{\bar{ar}} = \mathbb{E} \left\{ \frac{c_1 |h_{U_{1i}}|^2}{R_1(|h_{rr}|^2 + c_2)} \right\} = \frac{c_1 \lambda_{rr}}{R_1 \lambda_{U_{1i}} (1 + c_2 \lambda_{rr})}.$$

Similarly, the CDF $\mathcal{F}_{\gamma_{\bar{br}}}(1)$, $\mathcal{F}_{\gamma_{\bar{rb}}}(1)$, and $\mathcal{F}_{\gamma_{\bar{ra}}}(1)$ can be obtained and the expressions are omitted due to space limitation. The term $\Pr\{M, N : M < 1\}$ in (11) is evaluated as

$$\begin{aligned} \mathcal{F}_M(1) &= \Pr\{M \leq 1\} = 1 - \Pr\{\gamma_{\bar{ar}} > 1, \gamma_{\bar{rb}} > 1\} \\ \mathcal{F}_M(1) &= \mathcal{F}_{\gamma_{\bar{ar}}}(1) + \mathcal{F}_{\gamma_{\bar{rb}}}(1) - \mathcal{F}_{\gamma_{\bar{ar}}}(1) \mathcal{F}_{\gamma_{\bar{rb}}}(1). \quad (18) \end{aligned}$$

Similarly, $\Pr\{M, N : N < 1\}$ is evaluated as

$$\mathcal{F}_N(1) = \mathcal{F}_{\gamma_{\bar{ra}}}(1) + \mathcal{F}_{\gamma_{\bar{br}}}(1) - \mathcal{F}_{\gamma_{\bar{ra}}}(1) \mathcal{F}_{\gamma_{\bar{br}}}(1). \quad (19)$$

The last term $\Pr\{M < 1 \cap N < 1\}$ with dependent events M and N can be similarly derived as

$$\begin{aligned} \Pr\{M < 1 \cap N < 1\} &= \mathbb{E}_{Z_0} [\Pr\{M \leq 1 \mid Z=Z_0, N \leq 1 \mid Z=Z_0\}] \\ &= \mathbb{E}_{Z_0} [(\mathcal{F}_M(1) \mid Z=Z_0) (\mathcal{F}_N(1) \mid Z=Z_0)] \\ &= \mathcal{F}_{\gamma_{\bar{ra}}}(1) \mathcal{F}_{\gamma_{\bar{rb}}}(1) + (\mathcal{F}_{\gamma_{\bar{ra}}}(1) - \mathcal{F}_{\gamma_{\bar{ra}}}(1) \mathcal{F}_{\gamma_{\bar{rb}}}(1)) \\ &\quad \int_0^\infty \mathcal{F}_{\gamma_{\bar{ar}}}(1) \Big|_{Z=Z_0} \times f_Z(Z_0) dZ_0 + \\ &\quad (\mathcal{F}_{\gamma_{\bar{rb}}}(1) - \mathcal{F}_{\gamma_{\bar{ra}}}(1) \mathcal{F}_{\gamma_{\bar{rb}}}(1)) \int_0^\infty \mathcal{F}_{\gamma_{\bar{br}}}(1) \Big|_{Z=Z_0} \times f_Z(Z_0) dZ_0 \\ &\quad + (1 - \mathcal{F}_{\gamma_{\bar{ra}}}(1) - \mathcal{F}_{\gamma_{\bar{rb}}}(1) + \mathcal{F}_{\gamma_{\bar{ra}}}(1) \mathcal{F}_{\gamma_{\bar{rb}}}(1)) \\ &\quad \int_0^\infty (\mathcal{F}_{\gamma_{\bar{ar}}}(1) \mathcal{F}_{\gamma_{\bar{br}}}(1)) \Big|_{Z=Z_0} \times f_Z(Z_0) dZ_0. \quad (20) \end{aligned}$$

Integrating (20), the expression for $\mathcal{F}_{M \cap N}(1)$ can be obtained as (16), shown at the bottom of previous page. Combining (16), (18), and (19), we can obtain the closed-form expression for P_{out}^L defined in (11).

REFERENCES

- [1] Y. M. Khattabi and M. M. Matalgah, "Performance analysis of multiple-relay ad cooperative systems over rayleigh time-selective fading channels with imperfect channel estimation," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 427–434, Jan. 2016.
- [2] J. Harri, F. Filali, and C. Bonnet, "Mobility models for vehicular ad hoc networks: a survey and taxonomy," *IEEE Commun. Surveys Tutorials*, vol. 11, no. 4, pp. 19–41, Dec. 2009.
- [3] X. Zhou, T. A. Lamahewa, and P. Sadeghi, "Kalman filter-based channel estimation for amplify and forward relay communications," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, Nov. 2009, pp. 1498–1502.
- [4] K. S. Gomadam and S. A. Jafar, "Impact of mobility on cooperative communication," in *Proc. WCNC*, vol. 2, Apr. 2006, pp. 908–913.
- [5] P. Raut and P. K. Sharma, "Full-duplex multi-user two-way relay systems with optimal scheduling," in *2018 10th International Conference on Communication Systems Networks (COMSNETS)*, Jan 2018, pp. 443–446.
- [6] C. Li, B. Xia, S. Shao, Z. Chen, and Y. Tang, "Multi-user scheduling of the full-duplex enabled two-way relay systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1094–1106, Feb. 2017.
- [7] Y. Jeon, Y. T. Kim, M. Park, and I. Lee, "Opportunistic scheduling for multi-user two-way relay systems with physical network coding," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1290–1294, Apr. 2012.
- [8] L. Yang and M. S. Alouini, "Performance analysis of multiuser selection diversity," *IEEE Trans. Veh. Technol.*, vol. 55, no. 6, pp. 1848–1861, Nov. 2006.
- [9] P. Sharma and P. Garg, "Achieving high data rates through full duplex relaying in multicell environments," *Trans. Emerging Tel. Tech.*, vol. 27, no. 1, pp. 111–121, Jan. 2016.
- [10] M. Duarte, C. Dick, and A. Sabharwal, "Experiment-driven characterization of full-duplex wireless systems," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4296–4307, Dec. 2012.
- [11] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Elsevier/Academic Press, Amsterdam, 2007.
- [12] H. Yang, W. Meng, B. Li, and G. Wang, "Physical layer implementation of network coding in two-way relay networks," in *2012 IEEE International Conference on Communications (ICC)*, June 2012, pp. 671–675.
- [13] X. Liang, S. Jin, X. Gao, and K. K. Wong, "Outage performance for decode-and-forward two-way relay network with multiple interferers and noisy relay," *IEEE Trans. Commun.*, vol. 61, no. 2, pp. 521–531, Feb. 2013.
- [14] P. Liu and I. M. Kim, "Performance analysis of bidirectional communication protocols based on decode-and-forward relaying," *IEEE Trans. Commun.*, vol. 58, no. 9, pp. 2683–2696, Sept. 2010.

A Planar Four-Port Integrated UWB and NB Antenna System for CR in 3.1GHz to 10.6GHz

N. Anveshkumar
Teaching Fellow, ECE, VIT
Bhopal, India
nellaanvesh@gmail.com

Dr. A. S. Gandhi
Professor, ECE, VNIT
Nagpur, India
asgandhi@ece.vnit.ac.in

Abstract— This paper is aimed to present a planar four-port integrated UWB and narrowband (NB) antenna system for cognitive radio (CR) technology in the UWB 3.1GHz to 10.6GHz. This system consists of a UWB antenna for spectrum monitoring and three NB antennas for communication. These ultra wideband and narrowband antennas are incorporated on an FR-4 substrate having dimensions 28mm×31mm×1.6mm. The ultra wideband antenna, attached to port 1, is capable of monitoring the complete FCC unlicensed UWB spectrum 3.1GHz to 10.6GHz. The three NB antennas accomplish either single or dual bands to access the complete 3.1GHz to 10.6GHz band for communication. In particular, the first narrowband antenna, linked at port 2, attains a single band ranging from 8.26GHz to 11.16GHz. The second narrowband antenna, allied at port 3, also yields a single operating band ranges from 4.29GHz to 6GHz. The third NB antenna, associated with port 4, achieves a dual band behaviour starting from 3.06GHz to 4.49GHz and 5.97GHz to 8.35GHz. The coupling between the antennas is less than -17dB across the complete UWB. The proposed antenna system is fabricated and tested. It is observed that there is a good agreement between the simulated and measured results.

Keywords—Cognitive radio technology, coupling, integrated antenna structure, narrow band antenna, spectrum monitoring, ultra wideband antenna.

I. INTRODUCTION

In 2002, FCC has suggested various techniques to improve the spectrum usage ability [1]. Among them, one method is to use the unused channels or spectrum holes with the help of continuous monitoring of the radio environment. This model is well known as cognitive radio technology. In this process, the cognitive radio area is continuously sensed for unused channels. If any unused channel is found then the channel is utilized for any other wireless communication services. As the UWB 3.1GHz to 10.6GHz is completely unlicensed [2], we must employ CR technology to improve the spectrum usage ability. This technology monitors the radio environment with the help of ultra wideband antennas. It realizes communication with the help of narrowband antennas. In general, the narrowband antennas use frequency reconfiguration switching mechanisms to achieve many operating frequencies using a single antenna. However, the frequency reconfigurable narrowband antennas are having too many drawbacks [3-5]. Moreover, the existing systems [5] were able to perform one communication task at a time. So, to overcome the drawbacks of these reconfigurable mechanisms and to perform multiple

communication tasks at a time, the concept of multi-port integrated UWB and NB antenna systems is introduced in [5]. In this regard, various multi-port integrated UWB and NB antenna systems are proposed in [6, 7]. In [6], authors presented a three-port integrated UWB and NB antenna system. The proposed system incorporates one ultra wideband antenna for spectrum monitoring and two narrowband antennas for communication. This system can carry out a maximum of two communication tasks at a time due to two narrowband antennas. However, the two NB antennas were able to cover only 24% of total UWB spectrum for communication. In [7], authors presented a five-port integrated UWB and NB antenna system. The proposed system incorporates one UWB antenna for spectrum monitoring and four NB antennas for communication. This system can carry out a maximum of four communication tasks at a time due to four narrowband antennas. The four NB antennas are able to cover total 100% UWB spectrum. The UWB and NB antennas are printed on an FR-4 substrate of dimensions 40mm×36mm×1.6mm. Then, further attempt is carried out to reduce the five-port integrated UWB and NB antenna system dimensions by minimizing the UWB antenna, substrate dimensions and number of NB antennas. After maintaining a minimum isolation of 17dB, acceptable UWB and required NB return loss performances, the system dimensions are obtained as 28mm×31mm×1.6mm. Furthermore, the number of NB antennas to cover the required UWB is found to be three after increasing bandwidth of the first NB antenna and removing second NB antenna. In this paper, we discuss the planar four-port integrated UWB and narrowband antenna system for cognitive radio technology in the UWB 3.1GHz to 10.6GHz. This system consists of a UWB antenna for spectrum monitoring and three NB antennas for communication. This system can carry out a maximum of three communication tasks at a time due to three narrowband antennas.

II. PROPOSED FOUR-PORT INTEGRATED UWB AND NB ANTENNA SYSTEM

In the design of proposed antenna system, the five-port integrated UWB and NB antenna configuration [7] was considered as a basic structure and then further changes are made to obtain the proposed four-port integrated antenna system. The major modifications in the five-port antenna system to form proposed system include UWB antenna at P1

and the NB antenna at P2. The dimensions of remaining NB antennas at P4 and P5 are slightly changed and their positions are also altered to achieve compact size. In the UWB antenna design of proposed system, the radius is reduced till the lower cut-off frequency reaches slightly below 3.1GHz in all the operating conditions. This gives a radius of 5.6mm from 8.65mm. Moreover, it is shifted towards left side to get reduced dimensions for the entire structure. The first NB antenna at P2 adopts partial ground to obtain wide bandwidth. It also eliminates the use of NB antenna, which is at P3 in the five-port system. The ultra wideband and three narrowband antennas are incorporated on an FR-4 substrate having dimensions 28mm×31mm×1.6mm as shown in Fig. 1. This FR-4 substrate has a dielectric constant of 4.4 and a loss tangent of 0.019. The design and optimization of the proposed structure is carried out in FEM algorithm based software HFSS.

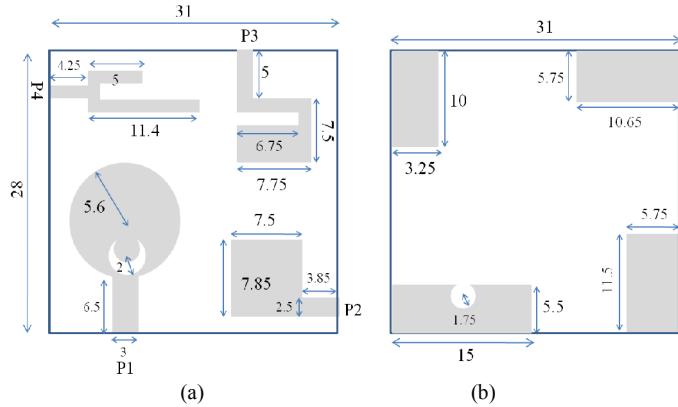


Fig. 1. Designed four-port integrated UWB and NB antenna structure. (a) Top view. (b) Bottom view [All dimensions are in mm].

In this system design, all the antennas are incorporated with partial grounds to achieve high impedance bandwidth. All the antenna dimensions are precisely optimized in the mentioned software. The fabricated model of the proposed antenna system is shown in Fig. 2.

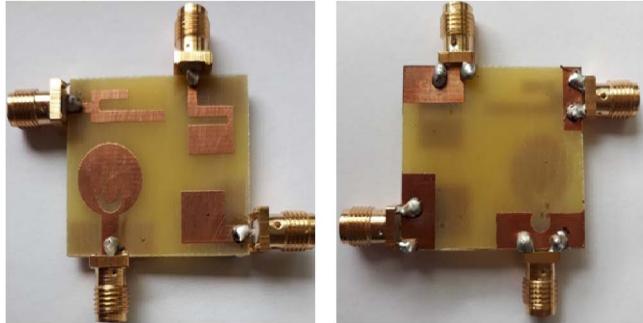


Fig. 2. Fabricated structure of the proposed system.

The experimentation shows that there is a good agreement between the simulated and measured results. But, there are shifts and changes due to fabrication problems, improper verification, impurities in materials, and losses in connectors. As pointed out in [5], in case of one unused channel identification and communication, the ultra wideband antenna and one narrowband antenna are used for spectrum monitoring and communication. The remaining narrowband antennas go on idle. In the proposed structure also, the ultra wideband

antenna and one narrowband antenna among the three narrowband antennas are selected. The remaining two NB antennas go on idle. The selection of one narrowband antenna is carried out by implementing excitation switching reconfiguration among the three NB antennas. However, the switching for selection of appropriate NB antenna relies on the identified unused channel frequency to which it matches. This complete process of switching the UWB and NB antennas forms three different operative cases. In one condition, the UWB antenna, linked at port 1, and one NB antenna, allied at port 2 (P2), are used for free channels monitoring and communication, respectively. The remaining NB antennas go on idle. In the other two cases also, the ultra wideband antenna and one narrowband antenna, linked at P3 or P4, are operated. The remaining NB antennas go on idle. During design analysis and testing of the proposed system, the antennas situated at idle are loaded with a 50Ω matching connector.

This paper presents the three operative cases in first, second, and third cases, respectively. It initially discusses performance analysis of the proposed structure in first operative case and then goes with second and third operative cases.

III. PERFORMANCE ANALYSIS IN THE FIRST OPERATIVE CASE

In this first operating case, the ultra wideband antenna, linked at port 1, and the narrowband antenna, allied at P2, are operated for free channels monitoring and communication, respectively. The remaining narrowband antennas associated at P3 and P4 are loaded with a 50Ω matching connector. The simulated and evaluated reflection coefficients of the reported ultra wideband antenna are depicted in Fig. 3. Basically, this UWB antenna is employed with partial ground to access ultra wideband.

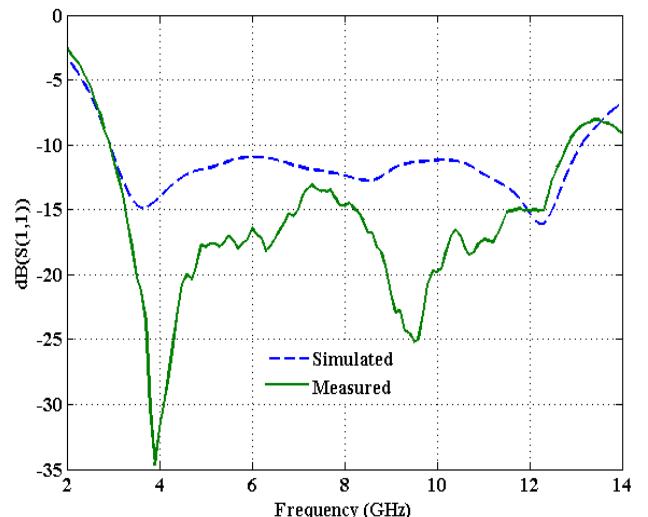


Fig. 3. Simulated and evaluated reflection coefficients of the reported ultra wideband antenna.

As shown in Fig. 3, it is clear that the ultra wideband antenna operates in the band from 2.91GHz to 13.1GHz that also covers the complete FCC unlicensed spectrum 3.1GHz to 10.6GHz. The antenna peak gains at different frequencies are presented in Fig. 4.

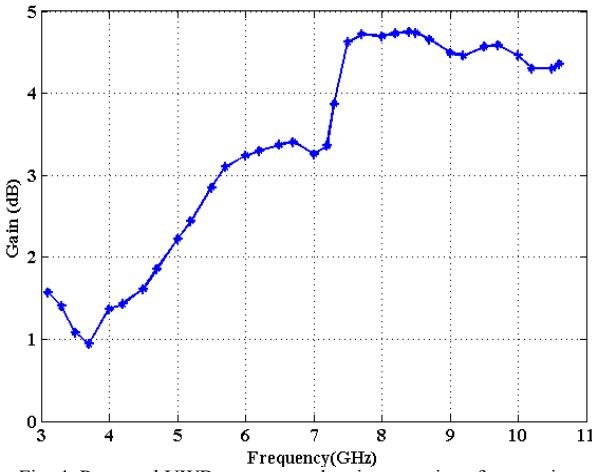


Fig. 4. Proposed UWB antenna peak gains at various frequencies.

From the Fig. 4 it is clear that as the frequency increases gain increases. This is because as the frequency increases directivity increases, which results to increase in gain [8]. The maximum peak gain is found to be 4.75dB at 8.4GHz. 2-D radiation pattern of this UWB antenna at the frequency of 8.4GHz is shown in Fig. 5.

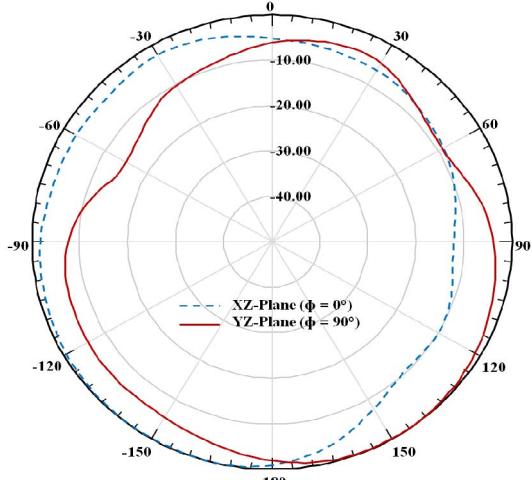


Fig. 5. Proposed UWB antenna 2-D radiation pattern at 8.4GHz.

From Fig. 5 it can be noted that the radiation pattern is almost omni directional. This is due to partial ground on the bottom plane [7]. The surface current on this UWB antenna at 8.4GHz is reported in Fig. 6.

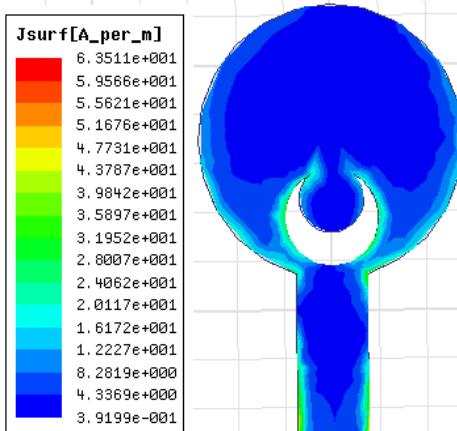


Fig. 6. Surface currents on the ultra wideband antenna at 8.4GHz.

The first narrowband antenna, allied at port 2, attains a single band ranging from 8.26GHz to 11.16GHz with a resonating frequency of 9.25GHz as shown in Fig. 7.

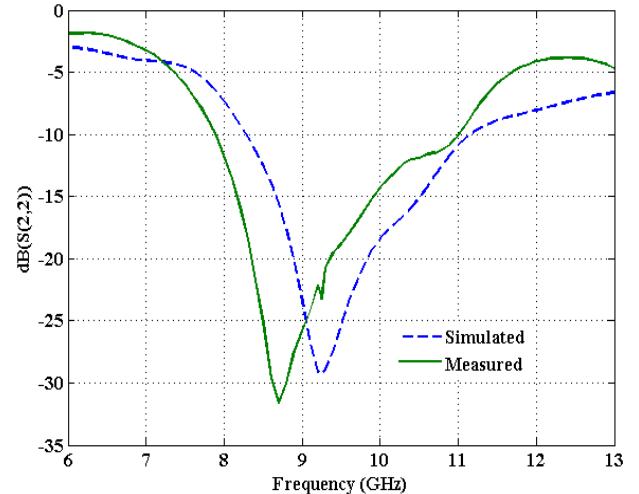


Fig. 7. Simulated and evaluated reflection coefficients of the reported first narrowband antenna.

This antenna is achieving a maximum gain of 2.63dB at 9.25GHz. 2-D radiation pattern of this NB antenna at the frequency of 9.25GHz is shown in Fig. 8.

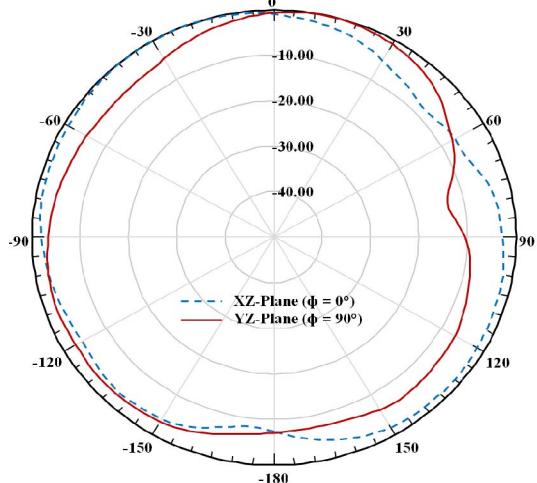


Fig. 8. Proposed narrowband antenna 2-D radiation pattern at 9.25GHz.

This antenna also yields omni directional radiation pattern due to partial ground. The surface currents on this NB antenna at 9.25GHz is reported in Fig. 9.

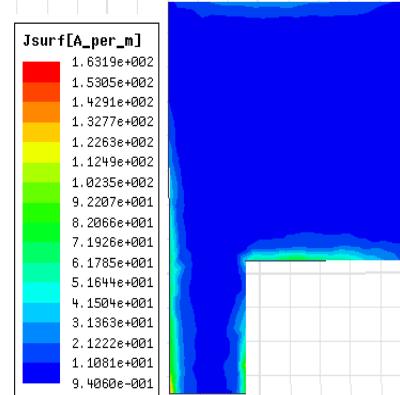


Fig. 9. Surface currents on the narrowband antenna at 9.25GHz.

The coupling between the ultra wideband and the first narrowband antennas is less than -17dB across the band 3.1GHz to 10.6GHz as shown in Fig. 10.

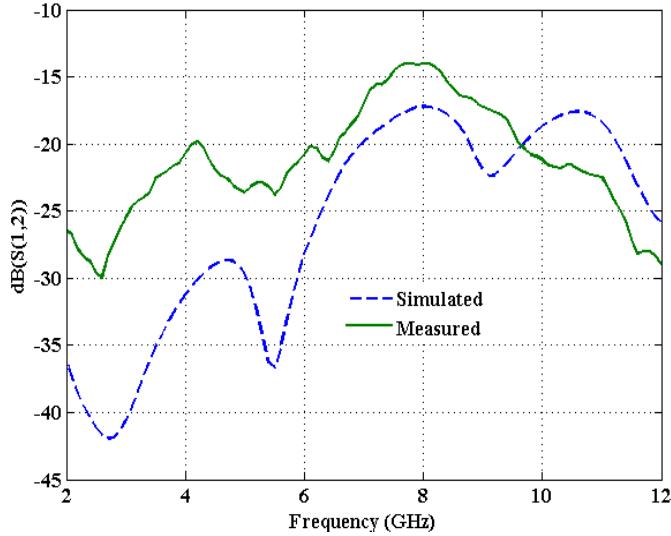


Fig. 10. Simulated and evaluated transmission coefficients between the ultra wideband and the first narrowband antennas.

IV. PERFORMANCE ANALYSIS IN THE SECOND OPERATIVE CASE

In this second operating case, the ultra wideband antenna, allied with port 1, and the narrowband antenna, associated at P3, are operated for free channels monitoring and communication, respectively. The remaining narrowband antennas associated at P2 and P4 are loaded with a 50Ω matching connector. The ultra wideband antenna achieves an operating band ranging from 2.94GHz to 13.15GHz as shown in Fig. 11.

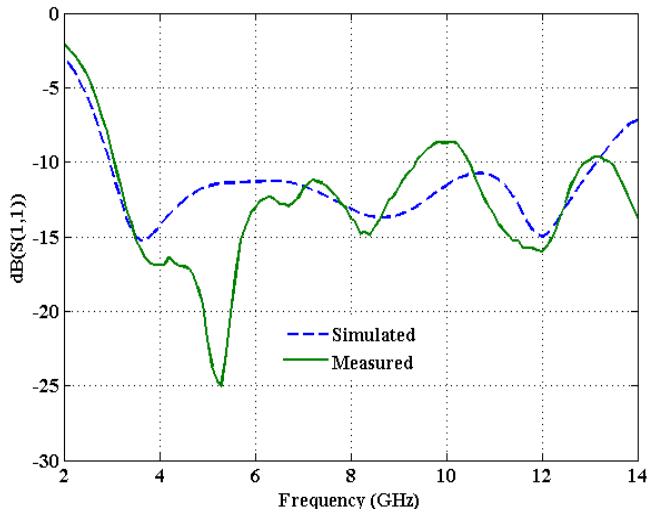


Fig. 11. Simulated and evaluated reflection coefficients of the reported ultra wideband antenna.

The second narrowband antenna, linked at port 3, also yields a single operating band ranges from 4.29GHz to 6GHz with a resonating frequency of 5.3GHz as shown in Fig. 12.

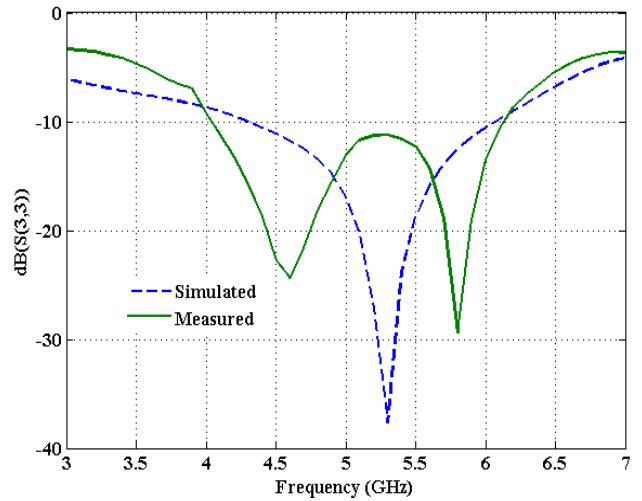


Fig. 12. Simulated and evaluated reflection coefficients of the reported second narrowband antenna.

The maximum gain is found to be 2.64dB at 5.3GHz. 2-D radiation pattern of this NB antenna at the frequency of 5.3GHz is shown in Fig. 13.

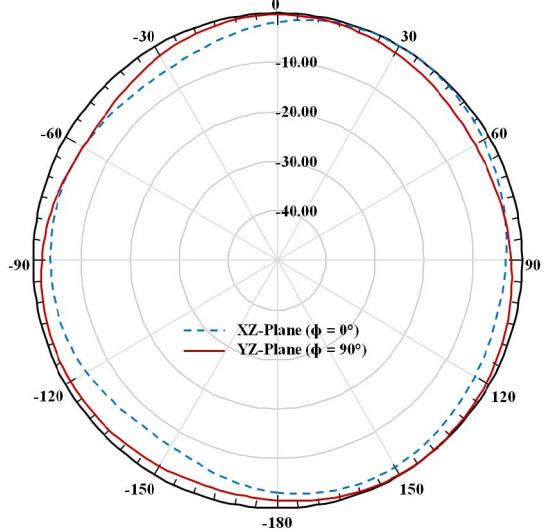


Fig. 13. Proposed narrowband antenna 2-D radiation pattern at 5.3GHz.

In Fig. 13 also the omni directional radiation pattern is observed. The surface currents on this NB antenna at 5.3GHz is reported in Fig. 14.

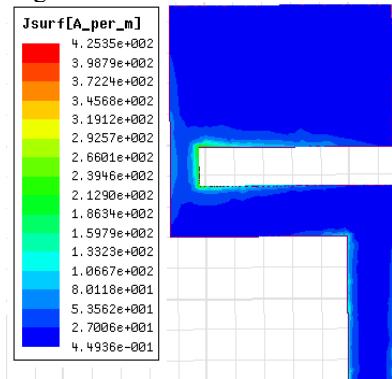


Fig. 14. Surface currents on the narrowband antenna at 5.3GHz.

The coupling between the ultra wideband and the second narrowband antennas is less than -23dB across the band 3.1GHz to 10.6GHz as shown in Fig. 15.

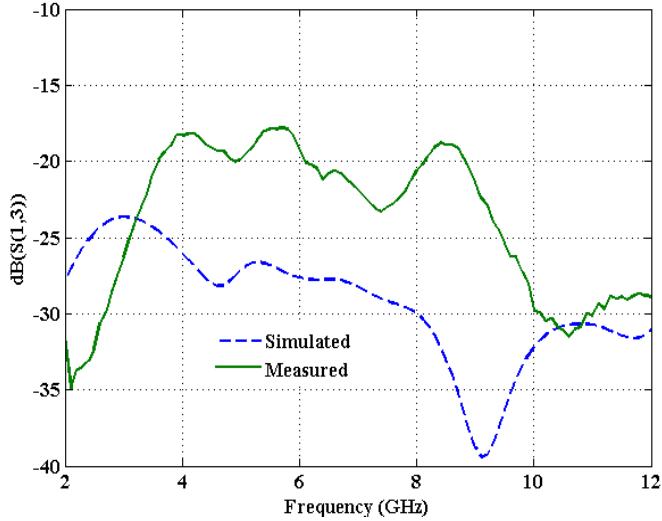


Fig. 15. Simulated and evaluated transmission coefficients between the ultra wideband and the second narrowband antennas.

V. PERFORMANCE ANALYSIS IN THE THIRD OPERATIVE CASE

In this third operating case, the ultra wideband antenna, allied with port 1, and the narrowband antenna, associated at P4, are operated for free channels monitoring and communication, respectively. The remaining narrowband antennas associated at P2 and P3 are loaded with a 50Ω matching connector. The ultra wideband antenna achieves an operating band covers from 2.47GHz to 13.64GHz as shown in Fig. 16.

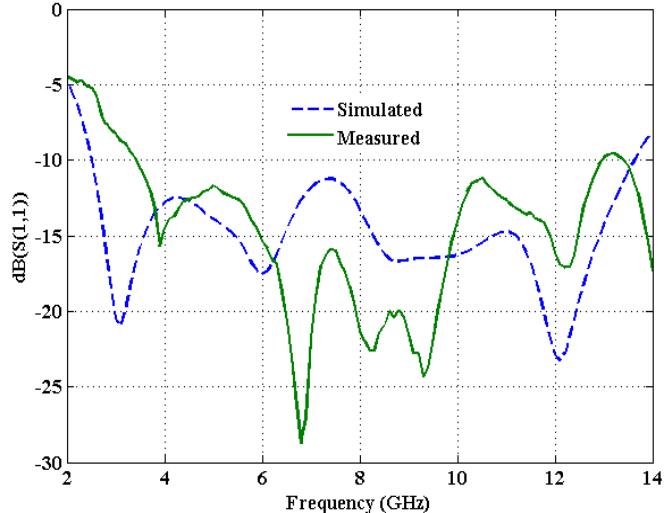


Fig. 16. Simulated and evaluated reflection coefficients of the reported ultra wideband antenna.

The third narrowband antenna, linked at port 4, achieves a dual band behaviour starting from 3.06GHz to 4.49GHz and 5.97GHz to 8.35GHz with resonant frequencies 3.85GHz and 6.55GHz, respectively as shown in Fig. 17.

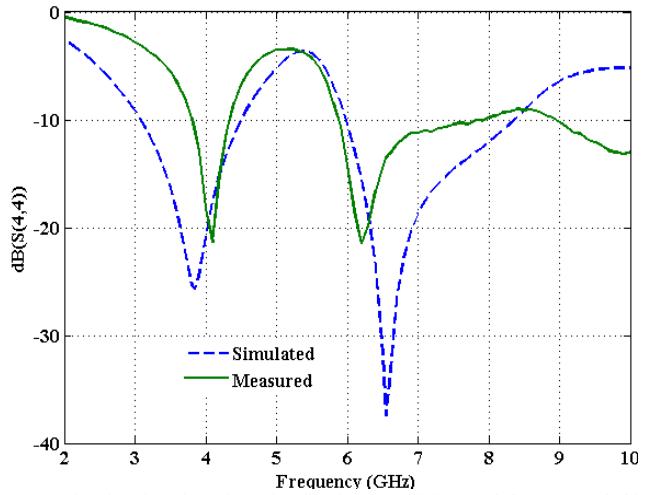


Fig. 17. Simulated and evaluated reflection coefficients of the reported third narrowband antenna.

The maximum gains are found to be 2.6dB and 2.53dB at 3.85GHz and 6.55GHz, respectively. 2-D radiation patterns of this NB antenna at the frequencies of 3.85GHz and 6.55GHz are shown in Fig. 18.

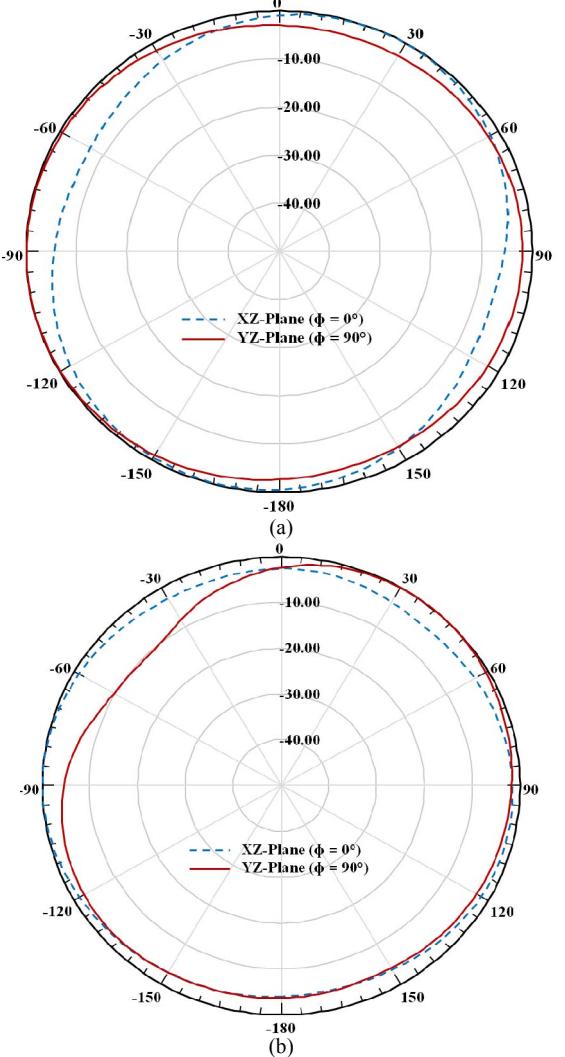


Fig. 18. Proposed narrowband antenna 2-D radiation patterns at (a) 3.85GHz and (b) 6.55GHz.

In Fig. 18 also the radiation pattern is found to be omni directional. The surface currents on this NB antenna at 3.85GHz and 6.55GHz are reported in Fig. 19.

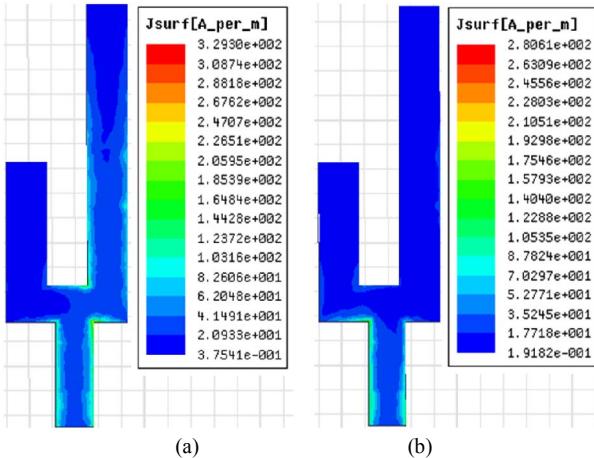


Fig. 19. Surface currents on the narrowband antenna at (a) 3.85GHz and (b) 6.55GHz.

From Fig. 19 it is clear that the strong currents are associated with longer patch at 3.85GHz and shorter patch at 6.55GHz. The coupling between the ultra wideband and the third narrowband antennas is less than -20dB across the band 3.1GHz to 10.6GHz as shown in Fig. 20.

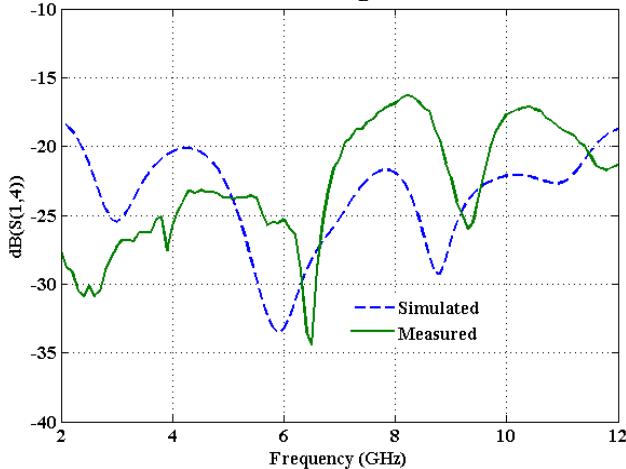


Fig. 20. Simulated and evaluated transmission coefficients between the ultra wideband and the third narrowband antennas.

VI. CONCLUSION

This paper presented a planar four-port integrated UWB and NB antenna system design for cognitive radio technology in the UWB 3.1GHz to 10.6GHz. In all the three operative cases, the ultra wideband antenna is capable of sensing the complete FCC unlicensed spectrum 3.1GHz to 10.6GHz. The three NB antennas achieve either single or dual bands to access the complete 3.1GHz to 10.6GHz band for communication. The coupling between the antennas was less than -17dB across the UWB, which is very well allowed. This structure at a time can perform a maximum of three communication tasks for enhancing the spectrum usage ability, which is the important goal of a cognitive radio model.

REFERENCES

- [1] Jayaweera S., Mosquera C., "A dynamic spectrum leasing (DSL) framework for spectrum sharing in cognitive radio networks", in Proceedings of the IEEE Forty-Third Asilomar Conference on Signals, Systems and Computers, pp. 1819-1823, Pacific Grove, California, Nov 2009.
- [2] FCC 1st report and order on Ultra-Wideband Technology, Feb. 2002.
- [3] Ahmed Rajae Raslan, "Metamaterial Antennas for Cognitive Radio Applications", A Thesis Submitted to the Electronics Engineering Department, American University in Cairo School of Sciences and Engineering in 2013.
- [4] Christos G. Christodoulou, Youssef Tawk, Steven A. Lane, and Scott R. Erwin, "Reconfigurable Antennas for Wireless and Space Applications", in Proceedings of the IEEE, Volume: 100, Issue: 7, pp. 2250 – 2261, April 2012.
- [5] Nella Anveshkumar, and Abhay Suresh Gandhi, "A Survey on Planar Antenna Designs for Cognitive Radio Applications", Wireless Personal Communications, Vol. 98, No. 1, pp. 541-569, Jan. 2018.
- [6] N. Anvesh Kumar, A. S. Gandhi, "A Compact Novel Three-Port Integrated Wide and Narrow Band Antenna system for Cognitive Radio Applications", International Journal of Antennas and Propagation, Volume 2016, Article ID: 2829357, 2016, 14 pages.
- [7] A. Nella, and A. S. Gandhi, "A Four-port Integrated UWB and Narrowband Antenna system Design for CR Applications", IEEE Transactions on Antennas and Propagation, Vol. 66, Issue. 4, pp. 1669-1676, February 2018.
- [8] N. Anvesh Kumar and A. S. Gandhi, "Small Size Planar Monopole Antenna for High Speed UWB Applications", in Proceedings of the Twenty Second National Conference on Communication (NCC), pp. 1-5, Guwahati, Assam, India, March 2016.

Generalized Selection Combining for Dynamic SSK-BPSK Systems

A. Ananth *, P. Maheswaran †, M. D. Selvaraj *

* Department of Electronics and Communication Engineering, IIITDM Kancheepuram, Tamil Nadu 600127.

ananth.iiitdm@gmail.com, selvaraj@iiitdm.ac.in

† Department of Electrical Engineering, IIT Madras, Tamil Nadu 600036.

mahswrn.iitm@gmail.com

Abstract—Space shift keying (SSK) is a multiple-input multiple-output (MIMO) technique in which the transmitter can be designed with a single radio frequency (RF) chain. By adaptively selecting the modulation in a two antenna transmitter as either SSK or binary phase shift keying (BPSK), dynamic SSK-BPSK (DSB) obtains second order transmit diversity. In this work, we conceive DSB with generalized selection combining (DSB-GSC) to reduce the receiver circuit complexity. Specifically, we propose the metrics of modulation selection and receiver antenna selection for DSB where the receiver is equipped with lesser number of RF chains than its antennas. The performance of DSB-GSC is analyzed with exact bit error rate (BER) expression which is validated using simulation results. From the results, we infer that DSB-GSC provides diversity order equal to twice the number of receiver antennas irrespective of the number of RF chains used at the receiver. We further infer that there is only small SNR gains attained for increasing receiver RF chains. Thus the receiver complexity of the system can be considerably reduced with a small performance loss compared to that of full complex receiver.

I. INTRODUCTION

In a single-input multiple-output (SIMO) wireless system, the adverse effects of fading is combated by diversity combining techniques in which N_r replicas of same signal received over many independently fading paths are combined in a specific manner [1]. Selection combining (SC) chooses the diversity branch that provides the best instantaneous signal-to-noise ratio (SNR), whereas maximal ratio combining (MRC) uses all N_r diversity branches with appropriate scaling to maximize the instantaneous SNR. Instead of using one best or all diversity branches, generalized selection combining (GSC) chooses N_c ($N_c \leq N_r$) best diversity branches and uses MRC among them [2].

Ever since the conception of wireless multiple-input multiple-output (MIMO) system, the prime focus of the researchers has always been in improving the spatial multiplexing and diversity gains by exploiting the properties of fading channels. Vertical Bell Labs layered space-time (V-BLAST) [3] employs multiple spatial streams of independent symbols for multiplexing gain. Whereas space time block code (STBC) [4] repeats spatial streams of independent symbols over space and time to achieve multiplexing as well as transmit diversity gain with a trade-off. Inherent to the transmitter and receiver

structure, the implementation of MIMO techniques (be it multiplexing or diversity) face difficulties. While implementing N_t antennas at the transmitter and N_r antennas at the receiver are usually inexpensive, the cost incurred in implementing radio frequency (RF) chains for those antennas is high in terms of power and circuitry [5].

For an $N_r \times N_t$ MIMO system, the complexity of both the transmitter and the receiver can be reduced by using fewer number of RF chains than antennas (at the transmitter/receiver) by using antenna selection (AS) techniques. The complexity reduction usually comes at the cost of slightly reduced performance [6], [7]. Antenna selection can be done for maximizing either the diversity gain or the multiplexing gain. Further, one can use antenna selection at the receiver, transmitter, or both. Implementing receiver antenna selection (RAS) is simpler as channel state information (CSI) is needed only at the receiver. But for transmitter antenna selection (TAS) and transmitter-receiver antenna selection (TRAS), the necessity of CSI at the transmitter further increases the complexity.

Spatial modulation (SM) is a recent MIMO multiplexing technique that takes a different approach to reduce complexity at the transmitter. In SM, only one RF chain is needed at the transmitter as it activates any one of N_t antennas with a conventional phase shift keying (PSK) or quadrature amplitude modulation (QAM) symbol to convey information through spatial constellation and signal constellation [8]. Due to the activation of a single transmitter antenna, SM avoids inter-channel interference (ICI) and inter-antenna synchronization (IAS) problems of conventional MIMO schemes [9]. Space shift keying (SSK) is a special case of SM which uses only spatial constellation (indices of transmit antennas) for information transmission [10]. The system performance of SM and SSK depend on the Euclidean distance among the random constellation points [8], [10].

A logical extension of the AS concept to SM and SSK is based on finding transmitter and/or receiver antennas that maximize the Euclidean distance between the random constellation points. A subset of antennas in the SM transmitter are selected based on CSI at the receiver in order to improve either the system performance [11] or the capacity [12]. The details about the selected antenna subset is fed back to the transmitter using feedback channel. In dynamic SSK-BPSK (DSB) [13], a one bit feedback from the receiver is used to adaptively switch

the modulation between binary SSK and a modified BPSK at the transmitter such that bit error rate (BER) performance is improved. For SSK in cooperative scenario, to select between direct path and relayed path, [14] uses the Euclidean distance between the SSK constellation points. Generalized Euclidean distance selection combining (GED-SC) for SSK is proposed in [15] where the receiver first calculates the minimum of $\binom{N_t}{2}$ correlated distances between SSK constellations for a receiver antenna, where (\cdot) denotes the binomial coefficient. Similar minimum is found for all N_r antennas. Out of the N_r minimum distances, those receiver antennas that give first N_c maximum distances where $N_c \leq N_r$ are chosen and MRC is used to combine the signal for detection.

Contribution: DSB proposed in [13] uses MRC with N_r RF chains at the receiver. In DSB, since both SSK and modified BPSK are used adaptively, conceiving a selection combining receiver with N_c RF chains where $N_c \leq N_r$ is of great interest as the selection metric for SSK is based on Euclidean distance whereas for modified BPSK, the selection metric is based on square magnitude of sum of the channels. Noting this, we conceive a GSC receiver with $N_c \leq N_r$ RF chains for DSB transmitter (abbreviated as DSB-GSC for simplicity in reference). We analyze the BER performance and diversity order of DSB-GSC in this work.

The paper is organized by conceiving the modulation selection and receiver antenna selection metrics of DSB-GSC in Section II. Further, Section II also discusses signal combining and detection at the DSB-GSC receiver. The exact BER performance of DSB-GSC is analyzed in Section III. Simulation results are given in Section IV to validate the BER analysis.

II. SYSTEM MODEL

The transmitter model of DSB as proposed in [13] is retained in DSB-GSC. At the DSB-GSC transmitter, two antennas with a single RF chain with output power constrained to ρ is considered. As in DSB, the DSB-GSC transmitter is assumed to be able to adaptively change its modulation between SSK and a modified form of BPSK (dubbed a-BPSK henceforth, signifying all-antenna BPSK) based on the feedback from the receiver. The difference between conventional BPSK and a-BPSK is that in the latter, a BPSK symbol is transmitted by simultaneously activating all two antennas, each with power $\rho/2$ [13] whereas conventional BPSK uses single transmitter antenna with power ρ to transmit a BPSK symbol. At the receiver of DSB-GSC, N_r antennas are considered. Further, it is considered that the receiver is equipped with N_c RF chains where $N_c \leq N_r$. The Rayleigh fading channel between the transmitter and the receiver is represented by $2 \times N_r$ matrix \mathbf{H} , where its independent and identically distributed (i.i.d.) entries $h_{i,j}$ at i th row and j th column are drawn from $\mathcal{CN}(0, 1)$ and \mathbf{h}_i signifies the i th column of \mathbf{H} . It is assumed that the receiver has the complete knowledge of \mathbf{H} . Based on this channel knowledge, DSB-GSC receiver firstly chooses whether transmitter should use SSK or a-BPSK and this decision is fed back to the transmitter. Secondly, based on the modulation selected, DSB-GSC receiver chooses N_c best

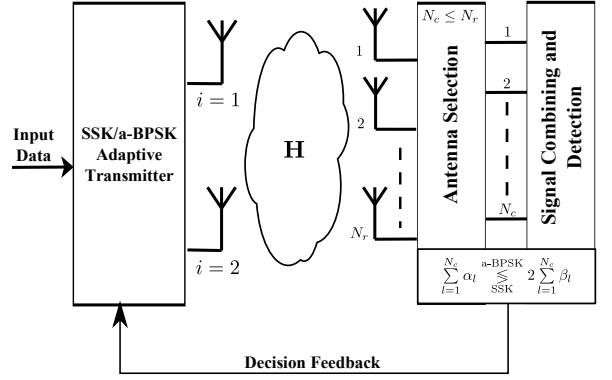


Fig. 1. Block Diagram of DSB-GSC

receiver antennas. Modulation and receiver antenna selection are done so as to improve the BER performance of DSB-GSC. These steps are depicted in Fig. 1 and are explained in the following section.

A. Modulation and Antenna Selection at the Receiver

From [10], it is known that the conditional BER of SSK depends on the Euclidean distance $\sum_{j=1}^{N_r} |h_{1,j} - h_{2,j}|^2$ between the SSK constellation points if MRC is used at the receiver with $N_c = N_r$ RF chains. Similarly for a-BPSK with MRC and N_r RF chains at the receiver, the conditional BER depends on $\sum_{j=1}^{N_r} |h_{1,j} + h_{2,j}|^2$ [13, Eq. (4)]. We define random variables $v_{1,j} \triangleq |h_{1,j} - h_{2,j}|^2$ and $v_{2,j} \triangleq |h_{1,j} + h_{2,j}|^2$.

Let $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_{N_r} \geq 0$ be the order statistics attained by arranging $\{v_{1,j}\}_{j=1}^{N_r}$ in the descending order of magnitude. Similarly, $\{\beta_l\}_{l=1}^{N_r}$ are obtained by arranging $\{v_{2,j}\}_{j=1}^{N_r}$ in the descending order. Since the receiver has complete knowledge of \mathbf{H} , it calculates α_l and β_l for $l \in \{1, \dots, N_r\}$ and computes

$$\sum_{l=1}^{N_c} \alpha_l \stackrel{\text{a-BPSK}}{\leq_{\text{SSK}}} 2 \sum_{l=1}^{N_c} \beta_l, \quad (1)$$

for modulation selection, where N can be chosen as N_c or as N_r (in which case modulation selection in DSB-GSC is same as that in DSB). Throughout the paper, we consider $N = N_c$. Once the decision from (1) is fed back to the transmitter, the DSB-GSC selects N_c best antennas (since the receiver is assumed to have N_c RF chains) to receive signal and perform MRC among them. The N_c receiver antenna indexes corresponding to the $\{\alpha_l\}_{l=1}^{N_c}$ are saved in set \mathcal{I}_1 . Similarly for $\{\beta_l\}_{l=1}^{N_c}$, the receiver antenna indexes are saved in \mathcal{I}_2 . Based on \mathcal{I}_1 and \mathcal{I}_2 , signal combining and detection are carried out as explained in the following section.

B. Signal Combining and Detection

Since DSB-GSC chooses N_c best antennas out of N_r , the received signal is modeled with $N_c \times 1$ vector. Let \mathbf{n} denote the additive white Gaussian noise (AWGN) of the receiver whose entries are i.i.d. random variables from $\mathcal{CN}(0, 1)$. When SSK is used at the transmitter, the signal received is given as

$\mathbf{y}_s = \sqrt{\rho} \tilde{\mathbf{h}}_a + \mathbf{n}$, where $a \in \{1, 2\}$ and $\tilde{\mathbf{h}}_a$ represents the N_c selected rows of channel vector \mathbf{h}_a based on the index set \mathcal{I}_1 . The transmitted symbol is detected after MRC as

$$\hat{a} = \arg \max_{a \in \{1, 2\}} \Re \left\{ \left(\mathbf{y}_s - \frac{\sqrt{\rho}}{2} \tilde{\mathbf{h}}_a \right)^H \tilde{\mathbf{h}}_a \right\}, \quad (2)$$

where in (2), $\Re\{c\}$ represents the real part of the complex number c and $(\cdot)^H$ gives the Hermitian transpose of a complex vector. When a-BPSK is used at the transmitter, the received signal is modeled as $\mathbf{y}_b = \sqrt{\frac{\rho}{2}} s_a \tilde{\mathbf{h}}^s + \mathbf{n}$, where we define $\mathbf{h}^s \triangleq \mathbf{h}_1 + \mathbf{h}_2$, $s_a \in \{-1, 1\}$ represents the BPSK symbol and $\tilde{\mathbf{h}}^s$ denotes the selected rows of \mathbf{h}^s using the indexes in \mathcal{I}_2 . Using MRC, the transmitted symbol is detected as

$$\Re \left\{ (\tilde{\mathbf{h}}^s)^H \mathbf{y}_b \right\} \stackrel{\hat{a}=2}{\gtrless} \stackrel{\hat{a}=1}{0}. \quad (3)$$

III. PERFORMANCE ANALYSIS

The decision to switch between a-BPSK and SSK is taken at the receiver based on (1). As MRC is performed at the receiver only among antennas from set \mathcal{I}_1 in case of SSK and \mathcal{I}_2 in case of a-BPSK, the instantaneous BERs are given as

$$P_{b_{SSK}}(\|\tilde{\mathbf{h}}_1 - \tilde{\mathbf{h}}_2\|) = Q \left(\sqrt{\frac{\rho}{2}} \|\tilde{\mathbf{h}}_1 - \tilde{\mathbf{h}}_2\|^2 \right), \quad (4)$$

$$P_{b_{a-BPSK}}(\|\tilde{\mathbf{h}}^s\|) = Q \left(\sqrt{\rho \|\tilde{\mathbf{h}}^s\|^2} \right). \quad (5)$$

It can be easily shown that $\|\tilde{\mathbf{h}}_1 - \tilde{\mathbf{h}}_2\|^2 = \sum_{l=1}^{N_c} \alpha_l$ and $\|\tilde{\mathbf{h}}^s\|^2 = \sum_{l=1}^{N_c} \beta_l$. Let us define the random variables $v_1 \triangleq \sum_{l=1}^{N_c} \alpha_l$ and $v_2 \triangleq \sum_{l=1}^{N_c} \beta_l$. DSB-GSC chooses SSK when $v_1 > 2v_2$ and a-BPSK otherwise. From this, the total error probability of DSB-GSC is given as

$$\begin{aligned} P_{b_{Total}} &= \int_0^\infty \int_0^{2v_2} Q(\sqrt{\rho v_2}) f_{V_1, V_2}(v_1, v_2) dv_1 dv_2 \\ &\quad + \int_0^\infty \int_0^{v_1/2} Q\left(\sqrt{\rho v_1/2}\right) f_{V_1, V_2}(v_1, v_2) dv_2 dv_1 \quad (6) \\ &= P_{b_{a-BPSK}} + P_{b_{SSK}}, \end{aligned} \quad (7)$$

where $P_{b_{a-BPSK}}$ and $P_{b_{SSK}}$ are the contributions of error due to a-BPSK and SSK in DSB-GSC, respectively. Since each v_{1,j_1} and v_{2,j_2} are statistically independent [13] for all j_1 and j_2 , the random variables v_1 and v_2 are also independently distributed. Substituting $f_{V_1, V_2}(v_1, v_2) = f_{V_1}(v_1)f_{V_2}(v_2)$ and using [16, Eq. (4.2)] for Gaussian Q -function in (6) leads to

$$\begin{aligned} P_{b_{Total}} &= \frac{1}{\pi} \int_0^{\pi/2} \left\{ \int_0^\infty \int_0^{2v_2} \exp\left(\frac{-\rho v_2}{2 \sin^2 \theta}\right) f_{V_1}(v_1) f_{V_2}(v_2) dv_1 dv_2 \right. \\ &\quad \left. + \int_0^{\infty} \int_0^{v_1/2} \exp\left(\frac{-\rho v_1}{4 \sin^2 \theta}\right) f_{V_1}(v_1) f_{V_2}(v_2) dv_2 dv_1 \right\} d\theta. \quad (8) \end{aligned}$$

For Rayleigh fading channel, $v_{1,j}$ and $v_{2,j}$ are exponentially distributed [16]. Since v_i for $i \in \{1, 2\}$ use sum of order statistics, their probability density function can be given as [16, Eq. (9.433)]

$$\begin{aligned} f_{V_i}(v_i) &= \binom{N_r}{N_c} \left\{ \frac{v_i^{N_c-1} \exp\left(-\frac{v_i}{2}\right)}{2^{N_c} (N_c-1)!} + \left(\frac{1}{2}\right)^{N_r-N_c} \sum_{n_1=1}^{N_r-N_c} (-1)^{n_1} \right. \\ &\quad \times \binom{N_r-N_c}{n_1} \left(\frac{-N_c}{n_1} \right)^{N_c-1} \exp\left(\frac{-v_i}{2}\right) \\ &\quad \times \left. \left[\exp\left(\frac{-n_1 v_i}{2 N_c}\right) - \sum_{m_1=0}^{N_c-2} \left(\frac{1}{m_1!} \right) \left(\frac{-n_1 v_i}{2 N_c} \right)^{m_1} \right] \right\}. \quad (9) \end{aligned}$$

Substituting (9) for $i \in \{1, 2\}$ in $P_{b_{a-BPSK}}$ given by (8), and integrating over the random variable V_1 using [17, Eq. (3.351.1)] gives

$$\begin{aligned} P_{b_{a-BPSK}} &= \frac{1}{\pi} \int_0^{\pi/2} \int_0^\infty \exp\left(\frac{-\rho v_2}{2 \sin^2 \theta}\right) \binom{N_r}{N_c}^2 \left\{ \left[\left(1 - \exp(-v_2) \right) \right. \right. \\ &\quad \times \sum_{k_1=0}^{N_c-1} \frac{v_2^{k_1}}{k_1!} \left. \right) + \sum_{n_1=1}^{N_r-N_c} (-1)^{N_c+n_1-1} \binom{N_r-N_c}{n_1} \\ &\quad \times \left(\frac{N_c}{n_1} \right)^{N_c-1} \left(\left(\frac{N_c}{N_c+n_1} \right) \left(1 - \exp\left(-v_2 \left(1 + \frac{n_1}{N_c} \right)\right) \right) \right. \\ &\quad \left. \left. - \sum_{m_1=0}^{N_c-2} \left(\frac{-n_1}{N_c} \right)^{m_1} \left(1 - \exp(-v_2) \sum_{k_2=0}^{m_1} \frac{v_2^{k_2}}{k_2!} \right) \right) \right] \\ &\quad \times \left\{ \frac{v_2^{N_c-1} \exp\left(-\frac{v_2}{2}\right)}{2^{N_c} (N_c-1)!} + \left(\frac{1}{2} \right)^{N_r-N_c} \sum_{n_2=1}^{N_r-N_c} (-1)^{(N_c+n_2-1)} \right. \\ &\quad \times \left(\frac{N_r-N_c}{n_2} \right) \left(\frac{N_c}{n_2} \right)^{N_c-1} \exp\left(\frac{-v_2}{2}\right) \left[\exp\left(\frac{-n_2 v_2}{2 N_c}\right) \right. \\ &\quad \left. \left. - \sum_{m_2=0}^{N_c-2} \left(\frac{1}{m_2!} \left(\frac{-n_2 v_2}{2 N_c} \right)^{m_2} \right) \right] \right\} dv_2 d\theta. \quad (10) \end{aligned}$$

Using [17, Eq. (3.351.3)], the infinite integral in (10) can be simplified to (11) given at the top of the next page, where $I_1(c_1, c_2)$ and $I_2(c_1, c_2, c_3, c_4)$ used in (11) are defined as

$$I_1(c_1, c_2) = \frac{1}{\pi} \int_0^{\pi/2} \left(\frac{\sin^2 \phi}{\sin^2 \phi + c_2} \right)^{c_1} d\phi, \quad (12)$$

$$I_2(c_1, c_2, c_3, c_4) = \frac{1}{\pi} \int_0^{\pi/2} \left(\frac{c_1 \sin^2 \phi}{c_2 \sin^2 \phi + c_3} \right)^{c_4} d\phi. \quad (13)$$

Similar to $P_{b_{a-BPSK}}$, $P_{b_{SSK}}$ (the second double integral term in (8)) can be found by simplifying the double integral on random variables V_1 and V_2 . The simplified expression of $P_{b_{SSK}}$ in terms of finite single integral is given by (14).

IV. SIMULATION RESULTS

The comparison of simulated performance of GSC in SSK, a-BPSK and DSB is given in Fig. 2 for $N_t = 2$ antenna

$$\begin{aligned}
P_{b_{a-BPSK}} = & \binom{N_r}{N_c}^2 \left\{ I_1(N_c, \rho) - \sum_{k_1=0}^{N_c-1} \frac{2^{k_1}(k_1+N_c-1)!}{k_1!(N_c-1)!} I_2(1, 3, \rho, N_c+k_1) + \sum_{n_1=1}^{N_r-N_c} (-1)^{n_1} \binom{N_r-N_c}{n_1} \left(\frac{-N_c}{n_1} \right)^{N_c-1} \left[\left(\frac{N_c}{N_c+n_1} \right) \right. \right. \\
& \times \left(I_1(N_c, \rho) - I_2(N_c, 3N_c+2n_1, \rho N_c, N_c) \right) - \sum_{m_1=0}^{N_c-2} \left(\frac{-n_1}{N_c} \right)^{m_1} \left(I_1(N_c, \rho) - \sum_{k_2=0}^{m_1} \frac{2^{k_2}(k_2+N_c-1)!}{k_2!(N_c-1)!} I_2(1, 3, \rho, N_c+k_2) \right) \Big] \\
& + \sum_{n_2=1}^{N_r-N_c} (-1)^{n_2} \binom{N_r-N_c}{n_2} \left(\frac{-N_c}{n_2} \right)^{N_c-1} \left[\left(I_2(N_c, N_c+n_2, \rho N_c, 1) - \sum_{k_1=0}^{N_c-1} 2^{k_1} I_2(N_c, 3N_c+n_2, \rho N_c, k_1+1) \right) \right. \\
& + \sum_{n_1=1}^{N_r-N_c} (-1)^{n_1} \binom{N_r-N_c}{n_1} \left(\frac{-N_c}{n_1} \right)^{N_c-1} \left(\left(\frac{N_c}{N_c+n_1} \right) \left(I_2(N_c, N_c+n_2, \rho N_c, 1) - I_2(N_c, 3N_c+2n_1+n_2, \rho N_c, 1) \right) \right. \\
& - \sum_{m_1=0}^{N_c-2} \left(\frac{-n_1}{N_c} \right)^{m_1} \left[I_2(N_c, N_c+n_2, \rho N_c, 1) - \sum_{k_2=0}^{m_1} 2^{k_2} I_2(N_c, 3N_c+n_2, \rho N_c, k_2+1) \right] \Big) \Big) - \sum_{m_2=0}^{N_c-2} \frac{1}{m_2!} \left(\frac{-n_2}{N_c} \right)^{m_2} \\
& \times \left(m_2! I_1(m_2+1, \rho) - \sum_{k_1=0}^{N_c-1} \frac{2^{k_1} (k_1+m_2)!}{3^{(k_1+m_2+1)} k_1!} I_1(k_1+m_2+1, \rho/3) + \sum_{n_1=1}^{N_r-N_c} (-1)^{n_1} \binom{N_r-N_c}{n_1} \left(\frac{-N_c}{n_1} \right)^{N_c-1} \right. \\
& \times \left[\left(\frac{m_2! N_c}{N_c+n_1} \left(I_1(m_2+1, \rho) - I_2(N_c, 3N_c+2n_1, \rho N_c, m_2+1) \right) \right) - \sum_{m_1=0}^{N_c-2} \left(\frac{-n_1}{N_c} \right)^{m_1} \left(m_2! I_1(m_2+1, \rho) \right. \right. \\
& \left. \left. - \sum_{k_2=0}^{m_1} \frac{2^{k_2} (k_2+m_2)!}{3^{(k_2+m_2+1)} k_2!} I_1(k_2+m_2+1, \rho/3) \right) \right] \Big) \Big] \Big\}. \tag{11}
\end{aligned}$$

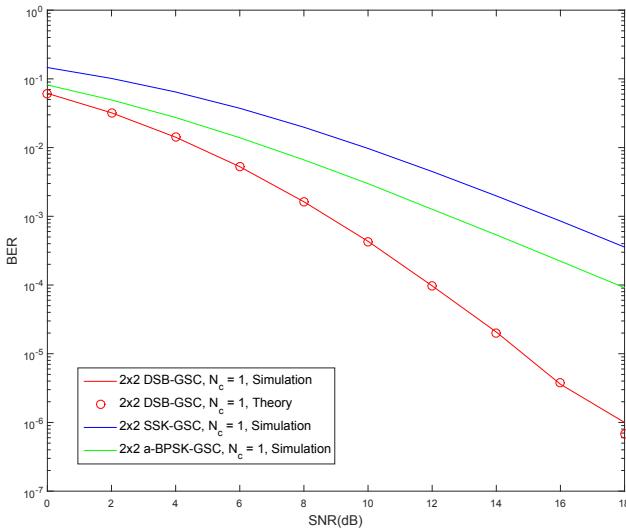


Fig. 2. BER vs. SNR (dB) for SSK, a-BPSK and DSB with GSC.

transmitter. Receiver combining and detection of symbols in SSK-GSC is based on (2), whereas (3) is used for the signal combining and detection in a-BPSK-GSC. For $N_r = 2$ antenna receiver with $N_c = 1$ RF chain, we observe that DSB-GSC performs better than a-BPSK-GSC and SSK-GSC. The performance improvement obtained in DSB-GSC is attributed to the adaptive switching of modulation at the transmitter. Further, in Fig. 2 we observe that the theoretical BER obtained by using (11) and (14) in (7) is validated by simulation, thus

verifying the BER analysis.

The influence of the number of receiver antennas on the diversity order of DSB-GSC is studied in Fig. 3. In DSB-GSC, the receiver first chooses the modulation to be used at the transmitter and then it chooses the receiver antennas for GSC. From [13] it is known that DSB achieves second order transmit diversity as it adaptively switches the modulation between SSK and a-BPSK. Further, GSC in a receiver with N_r antennas in SSK [15] and conventional MQAM/MPSK modulation [18] provides diversity gain of N_r . Since DSB-GSC uses both modulation selection and receiver antenna selection, it is natural that the diversity gain obtained in DSB-GSC is $2N_r$. For $N_c = 1$, the number of receiver antennas is varied from $N_r = 1$ to $N_r = 3$ in Fig. 3. For all the values of N_r , the theoretical and simulated BERs of DSB-GSC are plotted. To ascertain the diversity order obtained from these plots, $\frac{C_i}{SNR^{2i}}$ lines are also plotted where C_i are constants empirically calculated to closely follow the theoretical BER curves. From the plots in Fig. 3 we observe that $i = 1$ line runs parallel to $N_r = 1$ DSB-GSC system. Similar trends are observed in $i = 2, 3$ and $N_r = 2, 3$ curves in the high SNR region. From this, we infer that DSB-GSC achieves diversity order of $2N_r$ even with a single RF chain at the receiver. This reduces receiver circuit complexity without compromise on the diversity order.

To study the impact of varying number of receiver RF chains on the performance of DSB-GSC, Fig. 4 is plotted for $N_r = 3$. It can be observed from the plots in Fig. 4 that for various N_c values, the diversity order of the plots remain same at six since

$$\begin{aligned}
P_{b_{SSK}} = & \binom{N_r}{N_c}^2 \left\{ I_2(2, 2, \rho, N_c) - \sum_{k_1=0}^{N_c-1} \frac{2^{N_c}(N_c+k_1-1)!}{k_1!(N_c-1)!} I_2(1, 3, \rho, N_c+k_1) + \sum_{n_2=1}^{N_r-N_c} (-1)^{n_2} \binom{N_r-N_c}{n_2} \left(\frac{-N_c}{n_2}\right)^{N_c-1} \right. \\
& \times \left[\left(\frac{N_c}{N_c+n_2} \right) (I_2(2, 2, \rho, N_c) - I_2(2N_c, 3N_c+n_2, \rho N_c, N_c)) - \sum_{m_2=0}^{N_c-2} \left(\frac{-n_2}{N_c}\right)^{m_2} \left(I_2(2, 2, \rho, N_c) - \sum_{k_2=0}^{m_2} \left(\frac{2^{N_c}}{k_2!}\right) \right. \right. \\
& \times \left. \left. \frac{(N_c+k_2-1)!}{(N_c-1)!} I_2(1, 3, \rho, N_c+k_2) \right) \right] + \sum_{n_1=1}^{N_r-N_c} (-1)^{n_1} \binom{N_r-N_c}{n_1} \left(\frac{-N_c}{n_1}\right)^{N_c-1} \left[I_2(2N_c, 2(N_c+n_1), \rho N_c, 1) \right. \\
& - \sum_{k_1=0}^{N_c-1} 2I_2(N_c, 3N_c+2n_1, \rho N_c, k_1+1) + \sum_{n_2=1}^{N_r-N_c} (-1)^{n_2} \binom{N_r-N_c}{n_2} \left(\frac{-N_c}{n_2}\right)^{N_c-1} \left(\left(\frac{N_c}{N_c+n_2}\right) \right. \\
& \times (I_2(2N_c, 2(N_c+n_1), \rho N_c, 1) - I_2(2N_c, 3N_c+2n_1+n_2, \rho N_c, 1)) - \sum_{m_2=0}^{N_c-2} \left(\frac{-n_2}{N_c}\right)^{m_2} \left(I_2(2N_c, 2(N_c+n_1), \rho N_c, 1) \right. \\
& \left. \left. - \sum_{k_2=0}^{m_2} 2I_2(N_c, 3N_c+2n_1, \rho N_c, k_2+1) \right) \right) \right] - \sum_{n_1=1}^{N_r-N_c} \sum_{m_1=0}^{N_c-2} (-1)^{n_1} \binom{N_r-N_c}{n_1} \left(\frac{-N_c}{n_1}\right)^{N_c-1} \left(\frac{-n_1}{N_c}\right)^{m_1} \\
& \times \left[2^{m_1+1} \left(I_2(1, 2, \rho, m_1+1) - \sum_{k_1=0}^{N_c-1} \frac{(k_1+m_1)!}{m_1!k_1!} I_2(1, 3, \rho, k_1+m_1+1) \right) + \sum_{n_2=1}^{N_r-N_c} (-1)^{n_2} \binom{N_r-N_c}{n_2} \right. \\
& \times \left. \left(\frac{-N_c}{n_2}\right)^{N_c-1} \left(\left(\frac{N_c}{N_c+n_2}\right) (I_2(2, 2, \rho, m_1+1) - I_2(2N_c, 3N_c+n_2, \rho N_c, m_1+1)) - \sum_{m_2=0}^{N_c-2} \left(\frac{-n_2}{N_c}\right)^{m_2} \right. \right. \\
& \times \left. \left. \left(2^{m_1+1} \left(I_2(1, 2, \rho, m_1+1) - \sum_{k_2=0}^{m_2} \frac{(k_2+m_1)!}{m_1!k_2!} I_2(1, 3, \rho, k_2+m_1+1) \right) \right) \right) \right] \}. \quad (14)
\end{aligned}$$

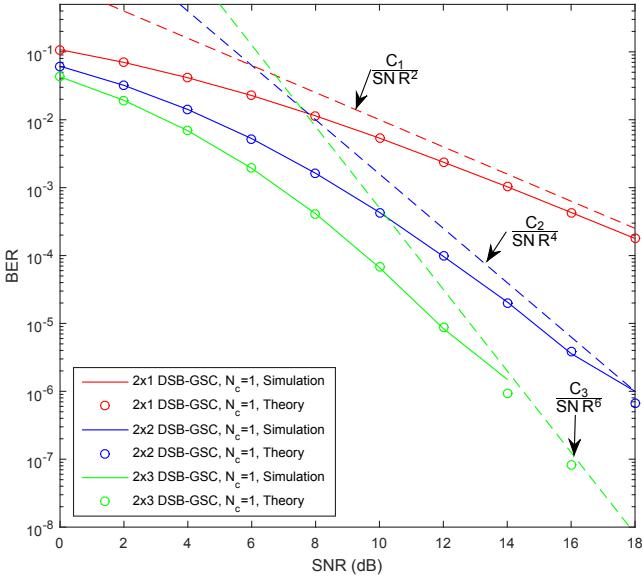


Fig. 3. Influence of N_r on the diversity order of DSB-GSC for a fixed N_c .

the curves run parallel to $\frac{C}{SNR^6}$. Compared to $N_c = 1$, the DSB-GSC receiver with $N_c = 2$ performs only 1.9 dB better and $N_c = 3$ performs 2.41 dB better at $BER = 10^{-3}$. From this we infer that the receiver complexity can be considerably reduced at the cost of a very small performance loss since

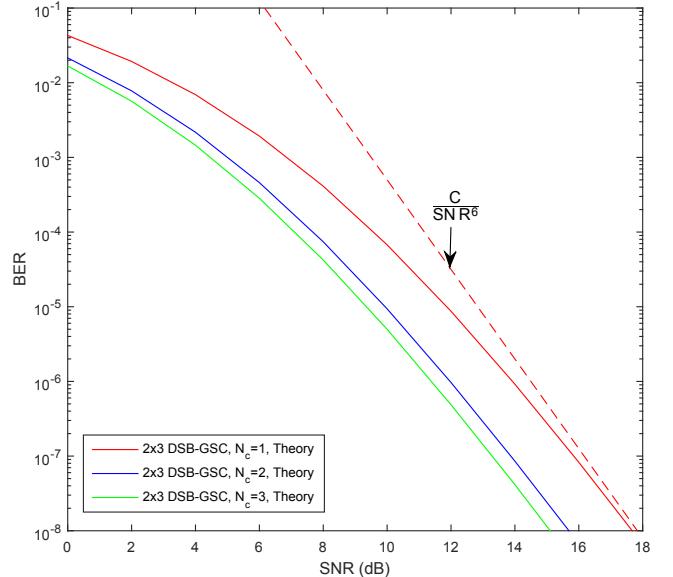


Fig. 4. Performance improvement in DSB-GSC for increasing N_c with $N_r = 3$. $N_c = N_r$ corresponds to ML detection with MRC.

in Fig. 4, we observe that the performance gain obtained by moving from $N_c = 2$ to $N_c = 3$ is only 0.51 dB at $BER = 10^{-3}$.

V. CONCLUSION

In this work we have introduced the system model of GSC in DSB. Specifically, the metrics of modulation selection and receiver antenna selection in DSB-GSC have been proposed. The performance of DSB-GSC has been analyzed by deriving the exact BER expressions in single finite integral form. Simulation results have been used to validate the analysis. From the results, it has been inferred that DSB-GSC provides better performance compared to GSC in SSK and BPSK since it uses modulation selection in addition to antenna selection at the receiver. Further, it has been inferred that DSB-GSC provides diversity order of twice the number of receiver antennas irrespective of the number of RF chains used at the receiver. For a given number of receiver antennas, it has been observed that for a linear increase in the RF chains at the receiver, there is only a small return in terms of SNR gain. Thus the receiver complexity can be considerably reduced with a small performance loss than that of full complex receiver.

ACKNOWLEDGMENT

This work was supported in part by the Department of Science and Technology - Science and Engineering Research Board (DST-SERB), Govt. of India (Ref. No. EEQ/2017/000243).

REFERENCES

- [1] D. G. Brennan, "Linear diversity combining techniques," *Proceedings of the IRE*, vol. 47, no. 6, pp. 1075–1102, Jun. 1959.
- [2] R. K. Mallik and M. Z. Win, "Analysis of hybrid selection/maximal-ratio combining in correlated Nakagami fading," *IEEE Transactions on Communications*, vol. 50, no. 8, pp. 1372–1383, Aug. 2002.
- [3] P. W. Wolniansky, G. J. Foschini, G. Golden, and R. A. Valenzuela, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Signals, Systems, and Electronics, 1998. ISSSE 98. 1998 URSI International Symposium on*. IEEE, 1998, pp. 295–300.
- [4] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.
- [5] J. H. Winters, "Smart antennas for wireless systems," *IEEE Personal Communications Magazine*, vol. 5, no. 1, pp. 23–27, Feb. 1998.
- [6] A. F. Molisch and M. Z. Win, "MIMO systems with antenna selection," *IEEE Microwave Magazine*, vol. 5, no. 1, pp. 46–56, Mar. 2004.
- [7] S. Sanaye and A. Nosratinia, "Antenna selection in MIMO systems," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 68–73, Oct. 2004.
- [8] J. Jeganathan, A. Ghayeb, and L. Szczecinski, "Spatial modulation: Optimal detection and performance analysis," *IEEE Communications Letters*, vol. 12, no. 8, pp. 545–547, Aug. 2008.
- [9] M. D. Renzo, H. Haas, and P. M. Grant, "Spatial modulation for multiple-antenna wireless systems: A survey," *IEEE Communications Magazine*, vol. 49, no. 12, pp. 182–191, Dec. 2011.
- [10] J. Jeganathan, A. Ghayeb, L. Szczecinski, and A. Ceron, "Space shift keying modulation for MIMO channels," *IEEE Transactions on Wireless Communications*, vol. 8, no. 7, pp. 3692–3703, Jul. 2009.
- [11] R. Rajashekhar, K. V. S. Hari, and L. Hanzo, "Quantifying the transmit diversity order of Euclidean distance based antenna selection in spatial modulation," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1434–1437, Sep. 2015.
- [12] R. Rajashekhar, K. V. S. Hari, K. Giridhar, and L. Hanzo, "Performance analysis of antenna selection algorithms in spatial modulation systems with imperfect CSIR," in *Proceedings of 19th European Wireless Conference*. VDE-Verlag, Apr. 2013, pp. 1–5.
- [13] P. Maheswaran and M. D. Selvaraj, "Performance analysis of feedback-based dynamic SSK-BPSK system," *IEEE Wireless Communications Letters*, vol. 5, no. 1, pp. 96–99, Feb. 2016.
- [14] P. Som and A. Chockalingam, "Bit error probability analysis of SSK in DF relaying with threshold-based best relay selection and selection combining," *IEEE Communications Letters*, vol. 18, no. 1, pp. 18–21, Jan. 2014.
- [15] A. Ananth and M. D. Selvaraj, "Error analysis of SSK with Euclidean distance based selection combining," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3195–3204, Apr. 2018.
- [16] M. K. Simon and M. S. Alouini, *Digital Communication over Fading Channels*, 2nd ed. Wiley, 2005.
- [17] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Academic Press, 2007.
- [18] M. Alouini and M. K. Simon, "An MGF-based performance analysis of generalized selection combining over Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 401–415, Mar. 2000.

Camera Zoom Detection and Classification Based on Application of Histogram Intersection and Kullback Leibler Divergence

Pavan Sandula, Manish Okade

Department of Electronics and Communication Engineering,

National Institute of Technology Rourkela,

Rourkela, India

516ec6004@nitrkl.ac.in, okadem@nitrkl.ac.in

Abstract—This paper presents a novel compressed domain technique for detecting zooming camera in video sequences and its further classification into zoom-in camera and zoom-out camera. The inter-frame block motion vector field serves as the input to the proposed system which is partitioned into four representative quadrants for analysis purposes. The histograms of these four quadrants are analyzed utilizing histogram intersection feature for zoom motion detection while the cumulative histogram of these four quadrants are analyzed utilizing Kullback-Leibler divergence feature for zoom motion classification purposes. Experimental validation carried out utilizing block motion vectors extracted using Exhaustive Search Motion Estimation algorithm as well as H.264 decoded block motion vectors demonstrate superior performance in comparison to existing techniques.

Index Terms—zoom motion, histogram intersection, Kullback-Leibler divergence, camera motion, compressed domain, support vector machine, block motion vectors.

I. INTRODUCTION

Motion in video sequences occurs due to either object motion, camera motion or due to combination of object as well as camera motions. The camera dynamics occur mainly due to the movement of the camera and needs to be recognized for video analysis purposes since it has various applications like autonomous navigation [1], video saliency estimation [2], video indexing and retrieval [3] to name a few. The motion of the camera can be translational wherein the motion is either in horizontal (referred as pan) direction or vertical (referred as tilt) direction or it can be zooming in nature where the environment under capture is brought near (referred as zoom-in) or taken away (referred as zoom-out) from the camera.

Due to the existence of various types of camera motion in video sequences namely panning, tilting, zooming etc. the first job at hand would be to detect zooming motion and later separate it into either zooming-in or zooming-out camera motion types which is the objective of the current work presented in this paper. Major focus on zoom motion in video sequences has been from the video coding domain

particularly the motion compensated prediction problem [4]–[6] for compression applications. However, zoom motion has also wide applications from video analysis point of view with applications ranging from indexing [7], retrieval [8], saliency estimation [9] to name a few. Zoom v/s non-zoom detection utilizing expectation maximization (EM) was carried out by Jin et al. [10]. However, since EM was utilized it had issues with initialization and convergence which affected the accuracy. Duan et al. [3] proposed an non parametric scheme for classifying camera motion categories with applications for video indexing and retrieval. They utilized mean shift clustering for identifying dominant motion clusters which was finally used for camera motion recognition. Since they used features namely cluster size, cluster number along with histograms of projected positions for identification purposes their method had shortcomings since these features were not able to bring out the underlying relationship of the block motion vectors. This method was improved in [11] wherein polar angle and magnitude histograms were used using a learning based scheme for identifying six camera motion types including zooming camera. However, in their work the zoom motion classification into zoom-in and zoom-out was left as future work since their focus was on translational camera for video stabilization applications. In [12] a transferable belief parametric model was utilized for the camera motion recognition problem. However, it utilized the Motion2D software to carry out the initial estimation of parameters, thereby not making it a stand alone algorithmic entity. In this paper, both the zoom motion detection as well as its further classification into zoom-in and zoom-out is carried out utilizing the concept of histogram intersection and Kullback-Leibler divergence [13] by analyzing the orientation histograms obtained by dividing the block motion vectors into four representative quadrants. Our results show superior performance in comparison to existing methods when tested using Exhaustive Search Motion Estimation (ESME) as well as H.264 compressed videos. Zoom motion detection and its classification into zoom-in/zoom-out plays a vital role in object localization which has applications in surveillance and autonomous navigation. Rest of the paper is organized as follows. Section II highlights

This research work is supported by SERB, Government of India under grant number: ECR/2016/000112.

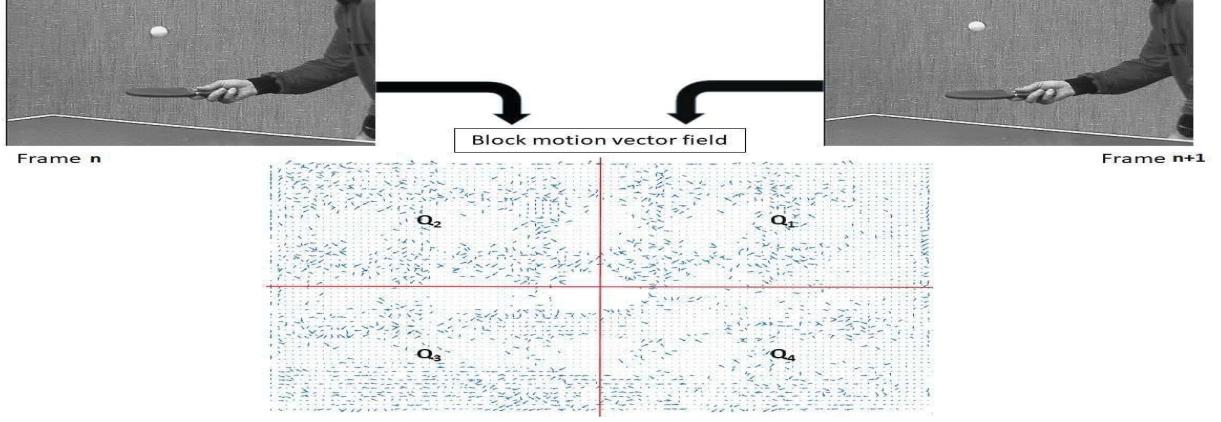


Fig. 1: Inter-frame Block Motion Vector Field.

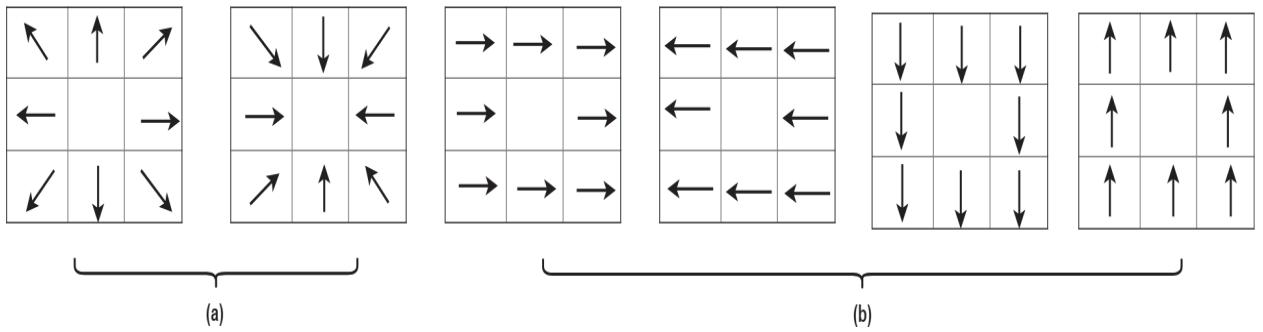


Fig. 2: Motion Vector Patterns corresponding to (a) zooming and (b) non-zooming camera.

the key contributions of the paper. Section III describes the proposed zoom motion detection and classification technique while Section IV gives the experimental results and finally in Section V we draw the conclusions.

II. KEY CONTRIBUTIONS

The contributions made in this work are two-fold. Firstly, zoom motion is recognized in video sequences i.e. identified from other camera motions like pan, tilt etc. utilizing the concept of histogram intersection between the quadrant histograms. Secondly, the identified zooming frames are further classified into zooming-in camera type and zooming-out camera type using the KL divergence between cumulative histogram of quadrants.

III. PROPOSED METHOD

A. Zoom motion detection

The proposed method utilizes the block motion vectors extracted from the compressed bitstream. Fig. 1 shows the inter-frame block motion vector field between two frames (frame # 33 and frame # 34) of sequence table tennis. As observed the block motion vector field shows various orientations corresponding to the nature of block motion vectors. The knowledge of nature of orientation pattern in case of zooming and non-zooming block motion vector fields will aid in detecting and separating the zooming camera from

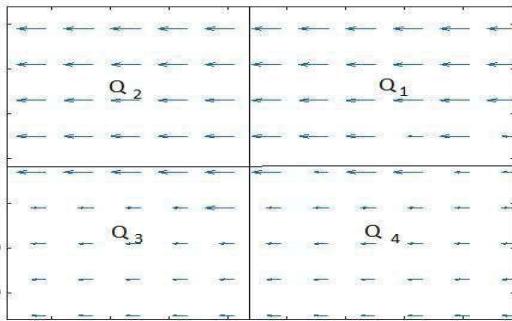
non-zooming camera. Fig. 2(a) shows the orientation patterns corresponding to the zooming camera motion pattern and Fig. 2(b) shows the orientation patterns corresponding to non-zooming category. As observed, the orientation pattern nature is different for the zooming camera and non-zooming camera which is exploited in this study by partitioning the block motion vector field into quadrants. The inter-frame block motion vector field between two frames is partitioned into 4 quadrants (Q_1, Q_2, Q_3, Q_4) for analysis purpose as shown in Fig. 4. The orientation histogram for the quadrants are estimated separately followed by calculating the histogram intersection between the quadrants to arrive at the feature vector which is utilized to train the C-SVM [14] classifier for separating the zooming frames from non-zooming frames. The detailed step by step description is given below;

- 1) Estimate the orientation of block motion vectors utilizing

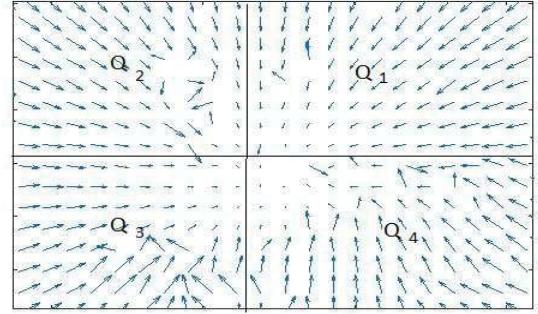
$$MV_{ori} = \arctan \left(\frac{MV^Y}{MV^X} \right) \quad 0 \leq MV_{ori} < 360 \quad (1)$$

where, $MV^Y \rightarrow$ vertical component of block motion vector and $MV^X \rightarrow$ horizontal component of block motion vector

- 2) Partition the inter-frame block motion vector field of size $N_1 \times N_2$ into 4 quadrants (Q_1, Q_2, Q_3, Q_4) as shown in



(a)



(b)

Fig. 3: Block motion vector field pattern for (a) Panning camera (b) Zooming camera.

Fig. 4, followed by estimating the orientation histogram of individual quadrants utilizing

$$H_{Q_i}(l) = \frac{1}{\frac{N_1}{4} \times \frac{N_2}{4}} \sum_{j=1}^R \sum_{k=1}^S f_1(Q_i(j, k); l) \quad (2)$$

where, $l \in [0^\circ, 360^\circ]$, $R \times S$ is size of individual quadrant (Q_i) with $R = \frac{N_1}{4}$ & $S = \frac{N_2}{4}$ and

$$f_1(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

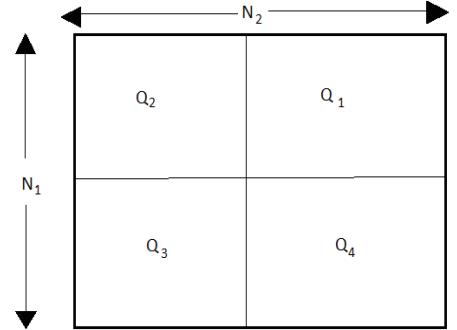
- 3) Estimate the histogram intersection (HI) between quadrants (Q_1 and Q_2), (Q_1 and Q_3), (Q_2 and Q_4) and (Q_3 and Q_4) utilizing

$$HI_{m,n} = \frac{\sum_{q=1}^l \min(H_{Q_m}(q), H_{Q_n}(q))}{\sum_{q=1}^l H_{Q_n}(q)} \quad (4)$$

where, $m \in (1, 1, 2, 3)$, $n \in (2, 3, 4, 4)$

- 4) Concatenate the histogram intersection of the quadrants estimated earlier to form the Feature Vector (FV)
- 5) Train the C-SVM classifier with linear kernel utilizing the feature vector formed for separating zooming frames from non-zooming frames.

The histogram intersection between two quadrants finds the amount of overlap between the orientation bins of the respective quadrants. If the orientation bins are similar for the quadrants under analysis as observed from Fig. 3 (a) then Eq. (4) tends towards 1 (i.e maximum overlap) signifying non-zoom motion (i.e pan, tilt). On the other hand for quadrants possessing dissimilar orientation bins as observed from Fig. 3 (b), Eq. (4) tends towards 0 (i.e least overlap) signifying zooming camera motion. This concept is exploited in the current work to distinguish between a zooming camera and a

Fig. 4: Inter-frame block motion vector field (size $N_1 \times N_2$) depicting the partitioning into 4 quadrants.

non-zooming camera. The rationale for computing histogram intersection between only specific pairs of quadrants is based on exploiting the concept of similarity between block motion vector orientations. As observed from Feature Vector formed in Eq. (5) adjacent quadrants (1, 2) & (3, 4) and diagonally opposite quadrants (1, 3) & (2, 4) are utilized to estimate the histogram intersection. In case of zooming frames these quadrants (adjacent and diagonal) result in least overlap due to dissimilar orientation types while in case of non-zooming frames (pan/tilt etc.) these quadrants result in maximum overlap due to similar orientation types.

B. Zoom motion classification

Once the zooming frames have been detected the next task would be to classify them into zooming-in camera and zooming-out camera. Fig. 5(a) and Fig. 5 (b) show the block motion vector field for zooming-in and zooming-out camera motion types, respectively. As observed, Zooming-in camera has motion vectors pointing outward from the center of frame

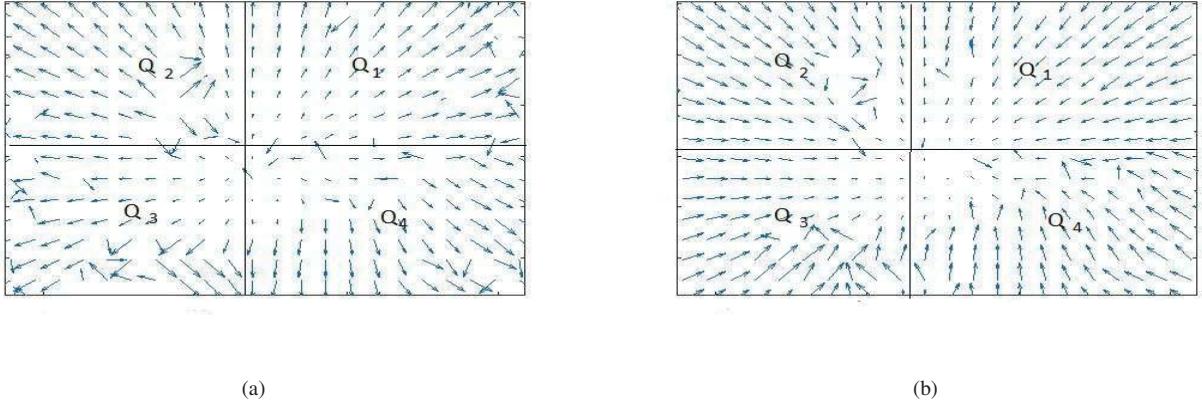


Fig. 5: Block motion vector field pattern for (a) Zooming-in camera (b) Zooming-out camera.

(i.e diverging field) while Zooming-out camera has motion vectors pointing towards the center of the field (i.e converging field). In-order to classify them, we utilize the Kullback-Leibler (KL) divergence between diagonal and adjacent quadrants as described below;

- 1) Estimate the orientation of block motion vectors utilizing Eq. (1).
- 2) Partition the inter-frame block motion vector field into 4 quadrants utilizing Eq. (2)
- 3) Estimate the cumulative histogram of individual quadrants utilizing

$$CH_{Q_i}(q) = \sum_{j=1}^q H_j \quad (6)$$

where, Q_i is the i^{th} quadrant with $i \in (1, 2, 3, 4)$ and $q \in (1 \dots 360)$

- 4) Estimate the Kullback-Leibler (KL) divergence between the cumulative histograms of quadrants (CH_{Q_1} and CH_{Q_3}), (CH_{Q_1} and CH_{Q_4}), (CH_{Q_2} and CH_{Q_3}), (CH_{Q_2} and CH_{Q_4}) utilizing

$$D_{KL}(CH_{Q_m} || CH_{Q_n}) = \sum_{q=1}^{360} CH_{Q_m}(q) \log \left(\frac{CH_{Q_m}(q)}{CH_{Q_n}(q)} \right) \quad (7)$$

where, $m \in (1, 1, 2, 2)$, $n \in (3, 4, 3, 4)$

- 5) Concatenate the KL-divergence of the quadrants to form the Feature Vector (FV)

$$\begin{aligned} FV = [& D_{KL}(CH_{Q_1} || CH_{Q_3}), D_{KL}(CH_{Q_1} || CH_{Q_4}), \\ & D_{KL}(CH_{Q_2} || CH_{Q_3}), D_{KL}(CH_{Q_2} || CH_{Q_4})] \end{aligned} \quad (8)$$

- 6) Train the C-SVM classifier with linear kernel utilizing the feature vector formed for separating zooming-in frames from zooming-out frames.

The KL divergence measures the amount of similarity between two distributions. In the current scenario, the cumulative histograms of quadrants are utilized to study the

behaviour of block motion vectors in case of Zooming-in camera and Zooming-out camera cases followed by estimating the KL-divergence between the cumulative histograms of the quadrants. Cumulative histogram (distribution) is chosen since it provides information of how the orientation patterns vary in each quadrant for zoom-in/zoom-out pattern types. We have not normalized the cumulative histogram before its application in Eq. (7) and plan to do it when we exploit the proposed zoom motion classification scheme for saliency application in future. In case of Zooming-in camera the KL divergence between the diagonally opposite and adjacent quadrants will be large since motion type is divergence (i.e vector pointing outwards) while for Zooming-out camera the KL divergence between diagonally opposite and adjacent quadrants will be relatively small since motion type is convergence (i.e vectors pointing inwards). This concept is utilized to separate the Zooming-in and Zooming-out camera motion types. The horizontally adjacent quadrants are excluded while estimating the KL divergence since it has been observed in our experimental simulation that including it in the feature vector does not significantly change the accuracy. This is due to the fact that whilst recognizing zoom-in v/s zoom-out camera types the maximum change in orientation will occur in vertically adjacent quadrants i.e. (1, 4) & (2, 3) and diagonally opposite quadrants i.e. (1, 3) & (2, 4).

IV. RESULTS

MATLAB R2016a is utilized for experimentation. Sequences available at <https://media.xiph.org/video/derf> and https://nsl.cs.sfu.ca/wiki/index.php/Video_Library_and_Tools which are standard in video analysis studies namely Tractor, Shields, Stefan, Station, Flowervase, Waterfall, Coastguard and Tempete are used. The zooming and non-zooming frames used in training and testing are manually labeled. Inter-frame block motion vectors are generated from these sequences by utilizing Exhaustive Search Motion Estimation (ESME) algorithm with block size '4×4', search range [-12 12] and cost function set to

TABLE I: Accuracy (%) for zoom motion detection at false positive rate set to 1%.

Block Motion Vector Type	Accuracy (%)		
	[3]	[11]	proposed method
ESME	91.08	92.25	96.71
ESME corrupted with gaussian noise ($\sigma^2 = 10$)	57.41	51.01	85.43
ESME corrupted with gaussian noise ($\sigma^2 = 20$)	51.25	50.16	70.00
ESME corrupted with gaussian noise ($\sigma^2 = 30$)	50.41	49.83	61.11
H.264	81.53	94.81	97.94

TABLE II: Area Under Curve (AUC) for zoom motion detection demonstrating the performance on various block motion vector types.

Block Motion Vector Type	proposed method
ESME	0.9958
ESME corrupted with gaussian noise ($\sigma^2 = 10$)	0.9289
ESME corrupted with gaussian noise ($\sigma^2 = 20$)	0.8297
ESME corrupted with gaussian noise ($\sigma^2 = 30$)	0.7338
H.264	0.9987

MAD. H.264/AVC obtained block motion vectors are also used to demonstrate the performance on a real codec by encoding them using JM19 encoder [15] (Software) with GOP IPP... and block size '4 × 4' to maintain consistency in comparison with ESME block size. Block motion vectors extracted from these encoded sequences (using Idecode.exe in JM-19) form the practical block motion vector case. Comparative studies is carried out with method proposed in [3] where dominant motion clusters were identified utilizing mean shift clustering followed by extracting features from the dominant clusters for camera motion recognition as well as method proposed in [11] where a learning based camera motion characterization scheme based on polar angle and magnitude histograms was utilized for recognizing six camera motion types.

A. Classifier Details

C-SVM with linear kernel is utilized for carrying out the classification studies. For each training and testing pair 40% of zoom and 40% of non-zoom samples are picked up randomly to train the C-SVM classifier and the remaining samples were used for testing. The above procedure is repeated thirty times using five fold cross validation on the training set. The cost parameter 'C' is trained and is used to obtain optimum cost in range $\{i | i \in \{0.1, 0.5..10\}\}$. 2000 frames from each class type are utilized for training the C-SVM and the frames from each class type which are not utilized for training are picked for testing. Same combination of sample size and classifier type is used in classifying zoom-in and zoom-out camera types. The detection accuracy is taken as the average of probability

of true positive rate (P_{tp}) and true negative rate (P_{tn}) and this is averaged over 30 random experiments utilizing

$$\text{Accuracy}(\%) = \left(\frac{P_{tp} + P_{tn}}{2} \right) \times 100 \quad (9)$$

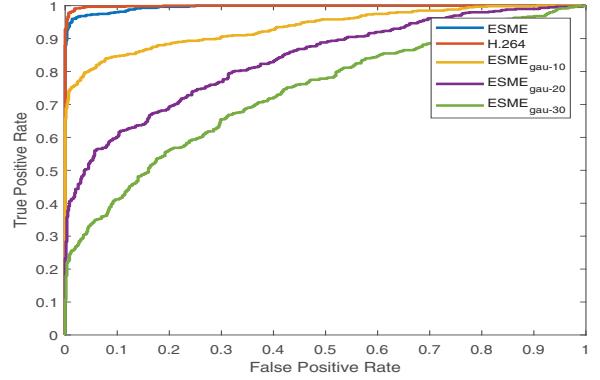


Fig. 6: ROC curves depicting the zoom detection (zoom v/s non-zoom) performance on various block motion vector types.

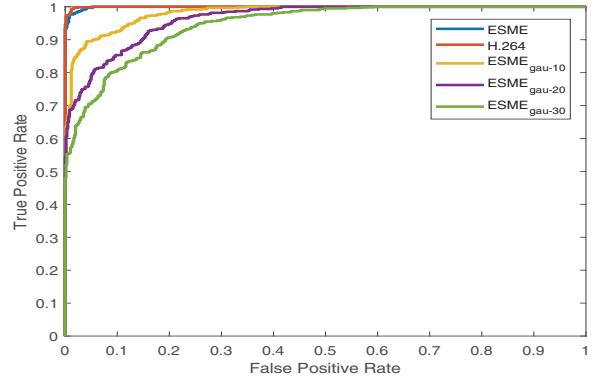


Fig. 7: ROC curves depicting the zoom classification (zoom-in v/s zoom-out) on various block motion vector types.

B. Objective Evaluation

Objective evaluation for the proposed technique is carried out in two ways: 1) ROC and AUC analysis which signifies the general detection performance and 2) by measuring detection accuracy at a low probability of false positives rate setting. The false positive rate is set to 1%, since this is the most widely used setting in classification studies. In order to analyze the robustness of the proposed method we add gaussian noise to both horizontal as well as vertical components of the block motion vector with zero mean and varying variance ($\sigma^2 = 10, 20, 30$) thereby generating 3 additional datasets for the experimental studies which we refer as $ESME_{gau-10}$, $ESME_{gau-20}$ and $ESME_{gau-30}$. Fig. 6 shows the ROC curves for ESME and its noise added variants and as observed

TABLE III: Accuracy (%) for zoom motion classification at false positive rate set to 1%.

Block Motion Vector Type	Accuracy (%)	
	[3]	proposed method
ESME	76.03	98.25
ESME corrupted with gaussian noise ($\sigma^2 = 10$)	63.53	89.24
ESME corrupted with gaussian noise ($\sigma^2 = 20$)	52.82	83.30
ESME corrupted with gaussian noise ($\sigma^2 = 30$)	50.50	76.56
H.264	60.31	98.55

TABLE IV: Area Under Curve (AUC) for zoom motion classification demonstrating the performance on various block motion vector types.

Block Motion Vector Type	proposed method
ESME	0.9991
ESME corrupted with gaussian noise ($\sigma^2 = 10$)	0.9822
ESME corrupted with gaussian noise ($\sigma^2 = 20$)	0.9636
ESME corrupted with gaussian noise ($\sigma^2 = 30$)	0.9468
H.264	0.9994

ESME achieves the best detection performance followed by drop on its noise added variants where it is noted that detection performance drops with increase in the variance of the added gaussian noise i.e. ($ESME_{gau-30} < ESME_{gau-20} < ESME_{gau-10}$). The corresponding AUC values are shown in Table II. The performance evaluation for Zoom motion classification is shown in Fig 7 which shows similar trend for ESME and its noise added variants. The corresponding AUC values are shown in Table IV.

Next, the detection accuracy at $FPR < 1\%$ obtained by the proposed method is shown in Table I and Table III. As observed the performance for the proposed method is better for all cases in comparison to existing methods thereby signifying the robustness of the proposed method which is due to the fact that the quadrant analysis using measures like histogram intersection for zoom motion detection and KL divergence for zoom motion classification is better able to capture the mutual relationship between the orientation of block motion vectors. It is observed from Fig. 6 and Fig. 7 that H.264 case achieves nearly same performance as ESME case in both zoom detection as well as zoom classification scenarios since H.264/AVC uses the concept of "skipped" motion inference wherein a skipped area of a predictively coded (P) frame infers motion content and aids in the detection as well as classification process which is very useful while coding video containing camera (global) motion.

V. CONCLUSIONS

This paper investigated the zoom motion detection as well as its further separation into zoom-in and zoom-out camera in case of compressed domain videos. The first motive was to detect zooming frames from non-zooming frames which was carried out utilizing the histogram intersection between quadrants as a feature. Once the zooming frames were detected, the next task was to separate them into zooming-in and zooming-out types which was carried out utilizing the KL divergence between quadrants as a feature. C-SVM classifier was utilized for training/testing purposes. Comparative analysis with existing methods using ESME as well as H.264 obtained block motion vectors showed very good performance for the proposed method. Our future work is focussed on exploiting the zooming cue for estimating salient regions in video sequences.

REFERENCES

- [1] S. Ghosh and J. Biswas, "Joint perception and planning for efficient obstacle avoidance using stereo vision," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept 2017, pp. 1026–1031.
- [2] Y. Fang, W. Lin, Z. Chen, C. Tsai, and C. Lin, "A video saliency detection model in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27–38, Jan 2014.
- [3] L.-Y. Duan, J. S. Jin, Q. Tian, and C.-S. Xu, "Nonparametric motion characterization for robust classification of camera motion patterns," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 323–340, 2006.
- [4] L.-M. Po, K.-M. Wong, K.-W. Cheung, and K.-H. Ng, "Subsampled block-matching for zoom motion compensated prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 11, pp. 1625–1637, 2010.
- [5] H.-S. Kim, J.-H. Lee, C.-K. Kim, and B.-G. Kim, "Zoom motion estimation using block-based fast local area scaling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1280–1291, 2012.
- [6] H. Yuan, Y. Chang, Z. Lu, and Y. Ma, "Model based motion vector predictor for zoom motion," *IEEE Signal Processing Letters*, vol. 17, no. 9, pp. 787–790, Sept 2010.
- [7] K. Schoeffmann, M. Taschwer, and L. Boeszoermenyi, "Video browsing using motion visualization," in *IEEE International Conference on Multimedia and Expo*, June 2009, pp. 1835–1836.
- [8] W. Pan and F. Deschenes, "Interpreting camera operations in the context of content-based video indexing and retrieval," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, June 2006, pp. 7–7.
- [9] G. Abdollahian, Z. Pizlo, and E. J. Delp, "A study on the effect of camera motion on human visual attention," in *15th IEEE International Conference on Image Processing*, Oct 2008, pp. 693–696.
- [10] R. Jin, Y. Qi, and A. Hauptmann, "A probabilistic model for camera zoom detection," in *16th IEEE International Conference on Pattern Recognition*, vol. 3. IEEE, 2002, pp. 859–862.
- [11] M. Okade, G. Patel, and P. K. Biswas, "Robust learning-based camera motion characterization scheme with applications to video stabilization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 453–466, 2016.
- [12] M. Guironnet, D. Pellerin, and M. Rombaut, "Camera motion classification based on transferable belief model," in *14th European Signal Processing Conference*, Sept 2006, pp. 1–5.
- [13] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 03 1951.
- [14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [15] *The H.264 AVC JM Reference Software*. [Online]. Available: <http://iphome.hhi.de/suehring/tm1/>.

Splitting Merged Characters of Kannada Benchmark Dataset using Simplified Paired-Valleys and L-Cut

H. R. Shiva Kumar, A. Madhavaraj and A. G. Ramakrishnan
Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, India
shivahr@gmail.com, madhavaraja@iisc.ac.in, agr@iisc.ac.in

Abstract—We reduce the computational complexity of the paired-valley algorithm for splitting merged characters, from $\Theta(N^2)$ down to $\Theta(N)$, where N is the number of symbols merged. We also propose an effective way (L-cut algorithm) to separate the merged half-consonants (known in Kannada as *ottus*) from the base symbols. We have created a benchmark dataset of 4033 sub-word images in Kannada, each comprising two or more merged characters. We test the recognition accuracy of Tesseract OCR on the created benchmark dataset, before and after applying our technique. The accuracy of Tesseract v3 OCR on the created dataset of 61.6% increases by 20% to a value of 81.7% after the splitting of the characters by our method. The algorithm's scalability to other scripts has been explored by limited experiments on Telugu and Tamil.

Index Terms—Merged characters, printed text, paired valleys, Kannada, ottu, Tamil, Telugu, OCR, VPP, Tesseract, old books, computational complexity.

I. INTRODUCTION

Merged characters that occur in old printed books are unseen patterns for the classifier and reduce the performance of optical character recognition (OCR) systems. They are akin to the occurrence of the out of vocabulary words in automated speech recognition systems. A number of techniques have been proposed to segment and recognize the merged characters, and can be basically classified as recognition-free and integrated segmentation-recognition (ISR) approaches. In the former approach, a set of rules is used to segment the characters before recognition, whereas in the latter approach, the recognition scores for the segmented components returned by the classifier are used to choose one out of many candidate segmentation paths.

A. Literature Survey

Zhu et al. [1] and Liu et al. [2] used ISR method to separate merged handwritten characters in Japanese and Chinese, respectively. Yang et al. [3] used the Vertical projection profile (VPP) of merged Chinese characters to obtain candidate cut locations (CCL) and chose the optimal ones using the recognition score. Messelodi and Modena [4] also split merged Roman characters using VPP. Davessar et al. [5] first vertically cut the merged Gurmukhi characters in the middle and searched for the CL within a window and confirmed it using recognition feedback and the aspect ratios of the separated units.

Bayer et al. [6] employed a statistical cut classifier and a search procedure to identify the merge locations in printed text, wherein, the computational complexity increases disproportionately with the number of merged characters in the image. Wang and Jean [7] deployed a neural network and shortest path to segment merged characters. Zhang, Tian and Li [8] use contour analysis to extract the concave points in the merged images of mathematical symbols and use them to postulate cut locations (CL) and employ a recognizer to verify them. Employing the histogram of the merged image to form binary-tree indexed demarcation using forward-backward algorithm, Tang et. al. [9] split the merged symbols.

Most of the recognition-free segmentation approaches proposed in the literature have dealt with merges in handwritten characters only. Congedo et al. [10] put forth the drop fall algorithm to split merged numeric strings in handwritten documents. Chang et al. [11] obtained the convex hull of the merged character and segmented merges in printed text using the features obtained from the concave residual and the shortest path algorithm. Lacerda and Mello [12] chose the optimal cut locations for splitting digit string merges in handwriting, employing the image skeleton and self-organizing maps.

Madhavaraj et. al. [13] reported the maiden effort in splitting of merged Kannada characters. They proposed a recognition based method for segmenting merged characters in printed Kannada documents by pairing top and bottom valleys of the merged-character image to locate the candidate cut locations (CCLs). Further, aspect ratios of the segmented parts, their recognition labels and scores returned by the classifier are used to choose the best segmentation path (SP). We refer to their algorithm as paired-valleys based ISR (PV-ISR).

B. Computational Complexity of PV-ISR method

A major computational complexity of the PV-ISR method is the need for exhaustive search for the optimal segmentation path from all possible paths that can be hypothesized from the candidate cut locations. For example, for the merged image shown in Fig. 1, the PV-ISR algorithm detects 7 CCLs, namely P1 to P7. The possible segmentation units (SU) are $\{B-P1, B-P2, \dots, B-E, P1-P2, P1-P3, \dots, P1-E, \dots, P7-E\}$ giving a total of 35 possible SUs, excluding input image B-E . Thus, the number of

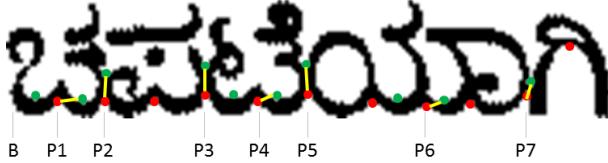


Fig. 1: Example of a Kannada sub-word with merged characters showing the candidate cut locations (CCLs) obtained from the PV-ISR algorithm. Top valley points are shown in green color and bottom ones, in red. Yellow lines show the CCLs, where there are matching pairs of top and bottom valleys in close proximity.

distinct SUs is $(N+1)(N+2)/2 - 1 = N(N+3)/2$, where N is the number of CCLs, and not $(N^2+3)/2$ as wrongly mentioned by Madhavaraj et al. in [13]. For each of those SUs, the PV-ISR algorithm extracts the required features and feeds it to a character recognizer. If the aspect ratio (AR) of any SU lies beyond the expected AR range of the recognized character class, then the recognition score of that SU is changed to a minimum value. Next, for finding the optimal SP, it considers the $(2^7 - 1) = 127$ possible SPs, computes the average likelihood of each of them, and finally selects the one with the maximum likelihood score. Further, the PV-ISR algorithm does not deal with merges of the base characters with the *ottus*, which types of merges also occur frequently in old Kannada texts.

C. Contributions of the paper

We simplify the paired-valleys algorithm for splitting of character merges so that it becomes a recognition-free approach and hence significantly reduces its computational complexity from $\Theta(N^2)$ down to $\Theta(N)$. The simplified algorithm can be applied as a pre-processing step before running any OCR, for the splitting of merged characters, as is demonstrated for the Tesseract OCR, in Sec. IV.

We also present a maiden algorithm for the detection and splitting (L-cut algorithm) of the merger of Kannada ottu symbols with the base characters.

For rigorously evaluating the performance of various algorithms, we have created a benchmarking dataset of 4033 Kannada sub-word images, each containing two or more merged symbols, along with their ground truth text in Unicode. Section IV compares the results of VPP-ISR, PV-ISR and our algorithm on this benchmarking dataset.

II. SIMPLIFIED PAIRED VALLEYS AND L-CUT (SPV-LC) ALGORITHM

A. Simplified Paired Valleys based Splitting

Merges in Kannada printed documents generally happen between the middle portions of successive characters, where they usually have outwardly rounded shape. This results in the formation of valleys just above and below the merged portion. Paired valleys based algorithms for splitting merged characters, such as PV-ISR [13], rely upon

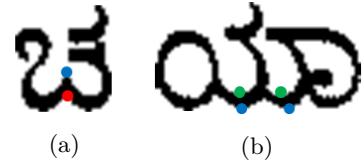


Fig. 2: Filtering out the valley points within a character: (a) Filter out bottom valleys (red color) in the ‘ω’ shaped portion of the character by looking for matching valleys (blue color) in the negative image. (b) Filter out top valleys (green color) in the ‘—’ shaped portions of the characters by looking for matching valleys (blue color) in the negative image at the bottom of the character.

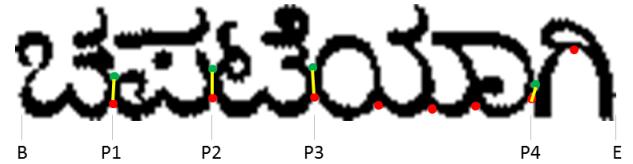


Fig. 3: Cut locations for the Kannada sub-word obtained by our SPV algorithm. Top valleys remaining after filtering are shown in green color and bottom ones, in red. Yellow lines show the cut locations, where there are matching pairs of top and bottom valleys in close proximity.

the detection of such pairs of valley points for determining the candidate cut locations (CCLs). Since valleys are also formed within a character, the CCLs could cut a valid character leading to oversegmentation as shown in Fig. 1 (at P1, P4 or P6). Here, each of the characters ಃ/ca/, ಣ/tte/ and ಯ್ಯ/yaa/ are cut into two pieces by the CCLs P1, P4 & P6, respectively. To avoid oversegmentation, recognition based approaches such as PV-ISR, consider all the possible combinations of CCLs, recognize the resulting segmentation units (SUs) and select the optimal sequence of CCLs that gives the maximum recognition score.

We propose a simple but effective approach to avoid oversegmentation by filtering out the valley points formed within a character, thereby solving the problem at the source. Characters in Kannada that have a ‘ω’ shaped portion at the bottom such as ಃ/ca/, ಃ/ja/, ಣ/tta/, ಃ/dda/, ಃ/da/ and ಃ/pa/ have a bottom valley at the mid of the character. Such bottom valleys among the background pixels have a counterpart among the foreground pixels immediately above them, as shown in Fig. 2a. Here, the bottom valley among the background pixels is shown in red color, and its counterpart among the foreground pixels, in blue. Similar pattern exists in characters that have an ‘m’ shaped portion at the top such as ಃ/nna/ or ಃ/i/, where the top valley among the foreground pixels has a counterpart among the foreground pixels immediately below it. Hence, to avoid oversegmentation, we filter out valley points that have an immediate counterpart among foreground pixels as shown in Fig. 2a.

Kannada characters that have \sim shaped portions at the bottom such as ಂ/ca/, ಃ/tta/ and ಯ/ya/ have top valleys at the mid of the \sim shape. Such top valleys among the background pixels also have their counterpart among the foreground pixels immediately below them at the bottom of the character as shown in Fig. 2b. Here, the top valleys among background pixels are shown in green color, and their counterparts among the foreground pixels, in blue. Again, we prevent oversegmentation by filtering out such top valleys possessing immediate counterparts among the foreground pixels at the bottom of the character, as shown in Fig. 2b.

Figure 3 shows the valleys detected by our algorithm on the same image shown in Fig. 1. Our algorithm has filtered out the valley points formed due to ω and \sim shaped portions, thereby avoiding oversegmentation.

B. Detecting and Splitting Base-Ottu Merges

Another common merge in Kannada characters occurs due to the touching of ottu symbols with the base symbols as shown in Fig. 4. Ottu symbols are additional graphemes used for representing consonant conjuncts and they appear below the *baseline* of the line [14]. Ottus can get merged with the previous consonant or the next one, leading to right-side or left-side merges, as shown in the first and second rows of Fig. 5. Some ottu symbols can appear right below the base consonant as shown in the third row of Fig. 5. We refer to this as the center merge.

An intuitive technique for splitting such merges is to cut the base-ottu merger just below the baseline and additionally along the left or right side of the base symbol, as shown in the second column of Fig. 5. We refer to this technique as L-cut (LC) due to the resemblance of the cut lines with the ‘L’ shape or its mirror image.

An important step before we can perform L-cut is the detection of baseline, for which we can leverage horizontal projection profile (HPP), as shown in Fig. 7. We smooth the HPP with a Gaussian function to remove jitter, and then take its first derivative. The location of the minimum of HPP derivative in the lower half of the image is taken as the position of the baseline. The third column of Fig. 5 shows the split of the base from the ottu symbols using L-cut.

To handle the scenarios, where there are merges among base characters as well as base-ottu merges, as shown in the first column of Fig. 6, we first invoke SPV based splitting on the image region above baseline. This splits the merges among the base characters, as shown in the second and third columns of Fig. 6. Then, on each connected component in the SPV split image, we detect if it corresponds to a base-ottu merger by looking at the extent of region below the baseline. If a base-ottu merger is detected, we invoke L-cut as shown in the fourth column of Fig. 6. The fifth column shows the components, after the splitting by the combined SPV-LC algorithm.

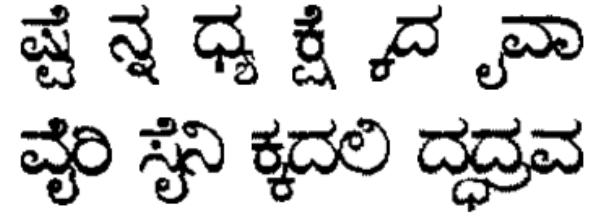


Fig. 4: Some example images, showing typical merges of the base characters (in the middle zone) with the ottu symbols (half consonants occurring below the baseline).

Input image	L-cut shown using green and red lines	Base-Ottus split using LC
ಪ್ರೇನ್	ಪ್ರೇನ್	ಪ್ರೇನ್
ದ್ವಾ	ದ್ವಾ	ದ್ವಾ
ಪ್ರೇನ್	ಪ್ರೇನ್	ಪ್ರೇನ್

Fig. 5: Some examples of base-ottu merges, separated (split) correctly using the proposed L-cut (LC) algorithm.

Input image	SPV splitting shown using yellow color	Symbols split using SPV	L-cut shown using green and red lines	Base-Ottus split using LC
ನೈ	ನೈ	ನೈ	ನೈ	ನೈ
ಜ್ವಾನಿ	ಜ್ವಾನಿ	ಜ್ವಾನಿ	ಜ್ವಾನಿ	ಜ್ವಾನಿ
ಸ್ವೀ	ಸ್ವೀ	ಸ್ವೀ	ಸ್ವೀ	ಸ್ವೀ

Fig. 6: Illustration of the SPV-LC algorithm, with 3 examples. First, the merged base characters are split by invoking the SPV algorithm on the image region above baseline. Then, the base-ottu merges are split using L-cut.

III. KANNADA DATASET FOR BENCHMARKING

For rigorously evaluating the performance of the proposed algorithm in splitting character merges, we have created a benchmarking dataset of 4033 sub-word level images, each containing two or more merged symbols. Figure 8 illustrates some of the images from this benchmarking dataset, showing merges across base components, as well as across base and ottu symbols. The ground truth text for each of the test images is provided in a separate Unicode text file as shown in Table I.

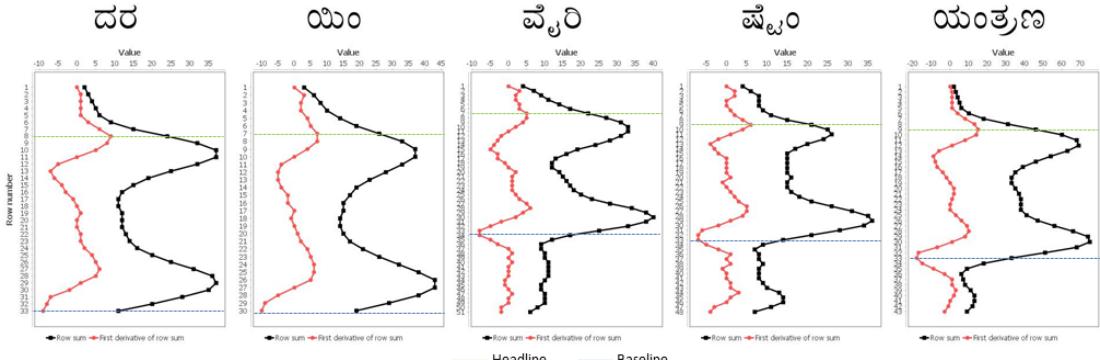


Fig. 7: Detecting headline and baseline using Gaussian smoothed horizontal projection profile and its derivative.

ಕೆಡು ದಪ್ಪ ಡಗ್ ಐಜ ಕೊ
ಫ್ರೆಗ್ ವ್ಯವಾ ಷ್ಟೈ ಸ್ನೈ ಜ್ಹಾನ್
ಚರವಾ ಎರಶಿ ಧ್ವವಾ ಜೆಟೊನಿ
ಕ್ರೆಡಲಿ ಯಂತ್ರಣಾ ಲ್ಯಾನ್ಪು ದ್ರದ್ರವ
ಇಳುತ್ತಿದ ಸಾಧ್ವಾಯಿ ನ್ಯಾಸ್ವನು
ಹರಣೆಗೆ, ಕ್ರೇವಾಡೆವೇ ದ್ರ ಧ್ವ ಪ್ಪು
ಪಾರಣೆಕ್ಕೊಗಿ ಫ್ಲಾಕ್ಸ್‌ನ್ಯೂಕ್ಕೆಡಿ

Fig. 8: Samples from the Kannada merged symbol dataset of 4033 sub-word images, each containing two or more merged symbols. Test images contain merges between base components, as well as between base and ottu components.

TABLE I: The ground truth text for each test image in the Kannada merged symbol dataset has been provided in a separate Unicode text file in the following format.

Image Name	Unicode Text	Image Name	Unicode Text
C0032.tif	ಕಟ್ಟ	C0223.tif	ಷ್ಟೈಂ
C0037.tif	ದಪ್ಪ	C0206.tif	ಸ್ನೈನಿ
C0033.tif	ಡಗ್	C0479.tif	ಜ್ಹಾನ್
C0228.tif	೦ ಐಜ	C0008.tif	ಚರವಾ
C0441.tif	ಇವಾ	C0125.tif	ಎರಶಿ
C0142.tif	ಫ್ರೆಗ್	C0199.tif	ಧ್ವವಾ
C0117.tif	ವ್ಯವಾ	C0053.tif	ಜೆಟೊನಿ

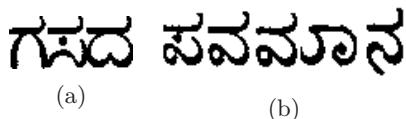


Fig. 9: (a) A sample Kannada sub-word, with merged symbols. (b) A letter press printed Kannada word.

Performance of the proposed algorithm is evaluated as the Levenshtein distance between the recognized and the ground truth texts, across the entire benchmarking dataset. Let N , S , I and D denote the number of Unicodes in ground-truth, substitutions, insertions and deletions, respectively. The Unicode level recognition accuracy is determined as:

$$\text{Accuracy} = (N - S - I - D)/N \quad (1)$$

This approach counts errors from both segmentation (merged symbols that are not split or those that are split incorrectly) and recognition phases. A better approach to measure errors only from the segmentation phase would require pixel-level ground truth of symbols, where all the pixels belonging to each symbol would be given an unique number in ground truth. For example, for sub-word ಗಸದ/gasada/ shown in Fig. 9a, all the pixels of the symbols ಗ/ga/, ಸ/sa/ and ದ/da/ would have values of 1, 2 and 3, respectively. In addition, for Indic scripts such as Kannada, such an approach should also standardize the list of OCR symbols for that script. For example, for the letter press printed Kannada word ಪವಮಾನ/pavamaana/, shown in Fig. 9b, the list of symbols could be {ಪ/pa/, ವ/va/, ಮಾ/maa/, ನ/na/} or {ಪ/pa/, ವ/va/, ವ/va/, ಏ/vowel_sign_aa/, ನ/na/}. However, the Unicode representation of that word is unique, consisting of Unicodes {ಪ/pa/, ವ/va/, ಮ/ma/, ಏ/vowel_sign_aa/ and ನ/na/}. Currently, the created benchmarking dataset has ground truth only at the Unicode/text level.

IV. RESULTS AND DISCUSSION

We compare the performance of our SPV-LC algorithm against those of the vertical projection profile (VPP-ISR) and paired-valleys (PV-ISR) based ISR algorithms on the new benchmarking dataset. The VPP-ISR and PV-ISR algorithms, as described in [13], are run on all the 4,033 images in the new benchmarking dataset, and the character recognition accuracy is computed as per eqn. (1). Table II shows the recognition accuracies of VPP-ISR and PV-ISR algorithms. For measuring the performance of our SPV-LC algorithm, we recognize the segmented

TABLE II: Comparison of the Unicode recognition accuracies (in %) of various splitting algorithms - VPP-ISR, PV-ISR and the proposed, on Kannada benchmarking dataset of 4,033 sub-word images, each containing 2 or more merged symbols. N, M: # of Unicodes in ground-truth and OCR output, respectively. N = 15,626. S, I and D: # of substitutions, insertions and deletions.

Algorithm	Accuracy	M	S	I	D
VPP-ISR	63.4	14,419	3,514	496	1,703
PV-ISR	81.0	14,160	1,365	70	1,536
SPV-LC	87.2	15,142	1,017	251	735

characters using the same set of features and classifier as in [13]; the combination of correlation and discrete wavelet transform is used as features and, support vector machine (SVM) with linear kernel is used as classifier. The last row of Table II shows the recognition accuracy of our SPV-LC algorithm. The VPP-ISR, PV-ISR and the SPV-LC algorithms achieve character recognition accuracies of 63.4%, 81% and 87.2%, respectively. The results show that the proposed algorithm achieves the best segmentation of the merged characters. Our results cannot be compared with those of [13], since (i) the database used there consists of complete words, whereas in our case, we use fully merged sub-words and (ii) the database used in [13] is not public.

The impact of L-cut is measured by running the SPV based splitting without using L-cut and measuring the difference in character recognition accuracy on the benchmarking dataset. Table III shows the recognition accuracy without using L-cut and the difference. The L-cut improves character recognition accuracy by 4.9%.

Since Tesseract OCR supports Kannada, we ran Tesseract v3 and v4 on each of the 4,033 sub-word images in the benchmarking dataset, and computed the character recognition accuracy as per (1). Table IV shows that Tesseract v3 achieves an accuracy of 61.6% on the test dataset, showing that this version of Tesseract has some in-built functionality for splitting character merges. However, Tesseract v4, which is based on LSTM (and hence, is good at capturing word-level language model), achieves a far lower accuracy of 29.9% on the sub-word level dataset, and hence we have skipped reporting of its results. Since our algorithm is recognition-free, it can be leveraged as an useful pre-processing step to split character merges before running the Tesseract OCR. Table IV also shows the performance of Tesseract v3 after preprocessing the input images using our SPV-LC algorithm. Accuracy of Tesseract v3 jumps from 61.6% to 81.7%, showing the utility of the proposed algorithm as a pre-processing step.

Figures 10a, 10b and 11 show examples of successful split of merged symbols using our SPV-LC technique. Figures 12a and 12b show examples, where our algorithm is either unable to split the merged symbols, or splits them

TABLE III: Impact of L-cut: recognition accuracy (in %) without using L-cut and the improvement, due to L-cut. All the notations are the same as in Table II. N = 15,626.

Algorithm	Accuracy	M	S	I	D
SPV without LC	82.3	14,509	1,266	194	1,311
LC improvement	+4.9	+633	-249	+57	-576

TABLE IV: Recognition accuracies (in %) of Tesseract OCR (v3.04.00) on the benchmark dataset, before and after splitting the input images using SPV-LC algorithm. All the notations are the same as in Table II. N=15,626.

	Accuracy	M	S	I	D
Tesseract v3	61.6	12,991	2,543	412	3,047
Tesseract v3 post SPV-LC	81.7	15,734	1,765	602	494

incorrectly. To see if the SPV-LC algorithm works for other Indic scripts, we tested it on a few merged symbols of Tamil and Telugu. Figures 13a and 13b show the outputs for these samples, which prove its effectiveness.

Input image	Symbols split using SPV-LC	Input image	Symbols split using SPV-LC
ಕೆಟ್	ಕೆ ಟ್	ಷ್ವೇಂದ್ರ	ಷ್ವೇ ಂದ್ರ
ದಪ್ತ	ದಪ್	ಷ್ವೇಂ	ಷ್ವೇಂ
ಡಗ್	ಡಗ್	ದ್ರು	ದ್ರು
೯ಜ್	೯ ಜ್	ಕ್ಷಾದಲ್	ಕ್ಷಾ ದಲ್

(a) (b)

Fig. 10: Some samples of merged symbols successfully split using SPV-LC technique. (a) Only base character merges. (b) Both base-base and base-ottu merges.

V. CONCLUSION

The SPV algorithm is successful in splitting merged symbols in a vast majority of cases, while also reducing the computational complexity to $\Theta(N)$ from $\Theta(N^2)$ in ISR approaches, where N is the number of symbols merged. The maiden L-cut algorithm is largely successful in detecting and splitting the merger of Kannada ottu symbols with the base characters. On the Kannada benchmark dataset, the Unicode recognition rate of Tesseract OCR increases from 61.6% to 81.7% after the splitting of the characters by our method. The algorithm's scalability to other scripts has also been explored by limited experiments on Telugu and Tamil. Thus, it holds promise as a useful module during character segmentation in existing OCRs. The standard, annotated database that has been created by us is now available for researchers [17].

REFERENCES

- [1] Zhu, B., Zhou, X. D., Liu, C. L., and Nagakawa, M., “A robust model for on-line handwritten Japanese text recognition,” Int. Jounal Document Analysis and Recog. 13(2), pp.121–131, 2010.
- [2] Liu, C. L., Jaeger, S. and Nakagawa, M., “Online recognition of Chinese characters: the state-of-the-art,” IEEE Trans. Pattern Analysis and Machine Intelligence, 24(2), pp.198–213, 2004.
- [3] Wuyi Yang, Shuwu Zhang, Haibo Zheng and Zhi Zeng, “A recognition-based method for segmentation of Chinese character in images and videos,” Proc. Int. Conf. Audio, Language and Image Processing, ICALIP July 2008, pp.723-728.
- [4] S. Messelodi and C.M. Modena, “Context driven text segmentation and recognition”, Pattern Recognition Letters, Vol. 17(1), pp. 47-56, Jan 1996.
- [5] Davessar, N.M., Madan, S. and Hardeep Singh, “A hybrid approach to character segmentation of Gurmukhi script characters,” Proc. 32nd Applied Imagery Pattern Recognition Workshop, Oct 2003, pp.169–173.
- [6] T. Bayer, U. Krebel and M. Hammelsbeck, “Segmenting merged characters”, Proc. XI Int. Conf. Pattern Recognition, Vol. II. Conf. B: Pattern Recognition Methodology and Systems, 1992.
- [7] Jin Wang and Jack Jean, “Segmentation of merged characters by neural networks and shortest path”, Pattern recognition, Elsevier, Volume 27, Issue 5, May 1994, pp. 649–658.
- [8] Dong-Yu Zhang, Xue-dong Tian and Xin-fu Li, “An improved method for segmentation of touching symbols in printed mathematical expressions,” Proc. 2nd Int. Conf. Adv. Computer Control (ICACC) March 2010, vol.2, pp. 251-25.
- [9] Tang Y, Li X, Zhang Y, Li M, Xu M, “Segmentation of touching characters via tree-indexed demarcation using forward and backward searches,” Adv. in Mech. Engineering, Oct 2017; 9(10).
- [10] Congedo, G., Dimauro, G., Impedovo, S. and Pirlo, G., “Segmentation of numeric strings,” Proc. Third Int. Conf. Document Analysis and Recognition, Aug 1995, vol.2, pp.1038-1041.
- [11] Chang, T.C. and Chen, S.Y, “Character segmentation using convex-hull techniques”. Int. J. Pattern Recognition and Artificial Intelligence, vol. 13, no. 6, pp. 833-858, 1999.
- [12] Lacerda, E.B. and Mello, C. A B, “Segmentation of touching handwritten digits using self-organizing maps,” Proc. 23rd IEEE Int. Conf. on Tools with Artificial Intelligence, 2011, pp.134-137.
- [13] Madhavaraj A., A. G. Ramakrishnan, Shiva Kumar H. R. and Nagaraj B., “Improved recognition of aged Kannada documents by effective segmentation of merged characters,” Proc. Tenth Int. Conf. Signal Processing and Communications, 2014.
- [14] Rituraj Kunwar, Shashi Kiran and A. G. Ramakrishnan. “Online handwritten Kannada word recognizer with unrestricted vocabulary,” Proc. Int. Conf. on Frontiers in Handwriting Recognition (ICFHR), IEEE, 2010, pp. 611–616.
- [15] Vijay Kumar B and A. G. Ramakrishnan. “Machine recognition of printed Kannada text.” International Workshop on Document Analysis Systems. Springer, Berlin, Heidelberg, 2002.
- [16] Vijay Kumar B and A. G. Ramakrishnan. “Radial basis function and subspace approach for printed Kannada text recognition.” IEEE International Conference on Acoustics, Speech, and Signal Processing 2004 May 17 (pp. V-321).
- [17] MILE, IISc “Kannada benchmarking dataset of merged symbols” <https://github.com/MILE-IISc/MergedSymbolsKannada> 2018.
- [18] H. R. Shiva Kumar and A. G. Ramakrishnan, “Gamma Enhanced Binarization - An adaptive nonlinear enhancement of degraded word images for improved recognition of split characters,” Proc. NCC 2019.

Input image	Symbols split using SPV-LC
ಘರದಶ್ವಕವಾಗಿ	ಘರದಶ್ವಕವಾಗಿ
ದಕ್ಷಿಂತೆ	ದಕ್ಷಿಂತೆ
ಹಾರುತಿರುವು	ಹಾರುತಿರುವು
ಕೆಟ್ಟಡಗಳೆ	ಕೆಟ್ಟಡಗಳೆ
ಹರಣೆಗೆ,	ಹರಣೆಗೆ,
ಕೈವಾಡುವೇ	ಕೈವಾಡುವೇ

Fig. 11: Some more examples of merged symbols successfully split by the SPV-LC technique.

Input image	Incorrect splitting by SPV-LC	Input image	Incorrect splitting by SPV-LC
ಗಂಡಲಿ	ಗಂ ದಲಿ	ಪ್ರೈ	ಪ್ರೈ
ಎರಿ	ಎರ ರಿ	ಪ್ರೈ	ಪ್ರೈ
ಕೋ	ಕ ಕೋ	ದಧ್ವವ	ದ ಧ್ವವ

(a)

(b)

Fig. 12: Inadequacy of SPV-LC algorithm. (a) Example images, where SPV technique failed or split the base symbols incorrectly. (b) Examples, where L-cut failed to correctly segment base-ottu merges.

Input image	Symbols split using paired-valleys technique	Input image	Symbols split using paired-valleys technique
தினா	தி னா	ரತ்வ	ರ த வ
லே	லೇ	இமா	இ மா
வேவி	வே வி	சீ	சீ
இரு	இ ரு	ಡீ	ಡೀ
தவி	த வி	ஓ	ஓ
கையி	கை யி	ரவ்ய	ர வ ய
றழுக	ற ழு க	கங்க	க ங க
ருக்குச	ரு கு கு ச	ದನ	ದ ந

(a)

(b)

Fig. 13: Testing of SPV-LC algorithm on other scripts. (a) Sample cases of successful splitting of Tamil merged symbols. (b) Splitting of merges between Telugu symbols.

Gamma Enhanced Binarization - An Adaptive Nonlinear Enhancement of Degraded Word Images for Improved Recognition of Split Characters

H. R. Shiva Kumar, and A. G. Ramakrishnan

Dept. of Electrical Engineering, Indian Institute of Science, Bangalore, India

shivahr@gmail.com, agr@iisc.ac.in

Abstract—Recognition performance of any OCR suffers because of the merged and split characters that occur in the scanned images of degraded printed documents. We propose an elegant method of non-linearly enhancing such degraded, gray-scale word images. This connects the broken strokes of the characters, so that binarization of the processed word images gives components with better connectivity for most characters or recognizable units. From an initial value of one, the value of gamma, the parameter determining the enhancement, is decreased in powers of 2 and the right value of gamma is chosen based on the recognition score of our character classifier. We have created a benchmark dataset of 1685 degraded word images obtained from scanned pages of several old Kannada books. The word images have been recognized before and after the proposed nonlinear enhancement. There is an absolute improvement of 14.8% in the Unicode level recognition accuracy of our SVM-based character classifier on the above dataset due to the proposed enhancement of the gray-scale word images. Even on the Google’s Tesseract OCR for Kannada, our gamma enhanced binarization results in an improvement of 5.6% in the Unicode level accuracy.

Index Terms—Split characters, printed text, power law transformation, gamma enhancement, binarization, Kannada, OCR, Tesseract, word images, old books.

I. INTRODUCTION

In the analysis and classification of images, irrespective of their domain of origin (medical, document, etc.), the key issue is one of reliable segmentation of the region of interest (tumor, abscess, word, characters, etc.) [1], [2]. The major challenges in dealing with the recognition of historical or severely degraded documents are: robust segmentation of characters in the presence of merged symbols and broken characters [3] and noise of different types. Approaches exist in the literature for addressing the issue of merged characters [4], [5]. This work addresses the issue of character splits caused during binarization due to poor printing or degradation of the paper due to aging. Figure 1 illustrates this problem with example word images from several Indic scripts. In this paper, we mainly consider split characters in Kannada script and test our proposed gamma enhanced binarization algorithm on a benchmark dataset created by us, and made publicly available [6].

Telugu	
Kannada	
Tamil	
Malayalam	
Gurumukhi	
English	

Fig. 1: Examples of degraded, binary, word images in various Indian languages, exhibiting splits (cuts) in characters.

A. Literature Survey

There have been different approaches in the literature to address the problem of recognition of broken characters. However, to the best of our knowledge, very few works have been reported on this problem for the Dravidian scripts, namely, Malayalam [7], Telugu [8], Kannada and Tamil. In fact, very few works have been reported in the literature on OCRs for Dravidian languages [9] such as Tamil [10]–[12] and Kannada [13], [14], [15]. Further, compound characters in Kannada have complex, two-dimensional arrangement [16], [17] and hence, segmentation of characters even from a good document is not a trivial task. Sachan et al. [18] have proposed a number of heuristic approaches, such as following the end points of cut strokes in the gray image, degree of overlap between connected components and the shortest distance between pairs of connected components (CC) to treat a set of CCs as belonging to a single character. Another obvious approach is to explore different combinations of consecutive connected components, recognize and then select the set of combinations that maximizes the recognition probability of a word. Following such an approach, Sumetphong and Tangwongsan [19] [20] treated this as a set-partitioning problem and propose a partition-growing

algorithm to group the broken pieces. The partition with the best posterior probability is chosen and the recognized output is corrected using a dictionary. Another possible approach is to start from the binarized symbols and apply morphological dilation and/or closing in an attempt to merge the broken pieces. Peerawit et al. [21] employed heuristics and morphological closing to reconstruct broken Thai characters only in the middle zone of the text line, whereas the cuts in the upper and lower zones were not considered. Yu and Yan [22] used conditional dilation, morphological analysis and stroke extension to reconstruct handwritten, broken digits.

Sulem and Sigelle [23] proposed the use of two, coupled dynamic Bayesian networks to model the interactions between the two streams of image columns and rows. They demonstrate a performance improvement of 1.1% over support vector machines (SVM) in recognizing broken characters in digitized Renaissance festival books. Droettboom [24] used graph combinatorics to join broken components of Roman characters from the Statistical Accounts of Scotland and evaluated the connected subgraphs using k-nearest neighbour classifier. Drira et al. [25] proposed a partial differential equation based formalism, combining the coherence-enhancing Weickert tensor driven diffusion filter and the singularities-preserving Perona-Malik scalar diffusion filter. The technique was applied on images from Gazette de Leyde to enhance the quality of degraded documents in Roman script and thus to reduce their OCR errors.

It is not an exaggeration, if we say that each of the above papers have used a different and custom database to report their results, which are not publicly available. Thus, in the absence of a common research database, there is no way one can compare the effectiveness, pros and cons of these different methods. In this context, it makes a difference to this field, if there are standard databases made publicly available for these kinds of degraded images, annotated with the corresponding ground truth.

B. Contributions of the paper

- Our proposed algorithm enhances the histogram of the gray level word image nonlinearly, in order to obtain the best binarized image for recognition.
- Though tested here only on Kannada word images, the proposed algorithm can be expected to work on any other script, provided the classifier in the OCR used provides a reliable recognition score for each recognized symbol.
- A standard, annotated database of 1685 degraded word images of Kannada script has been created, which has been made freely available [6] for benchmarking various algorithms against one another.

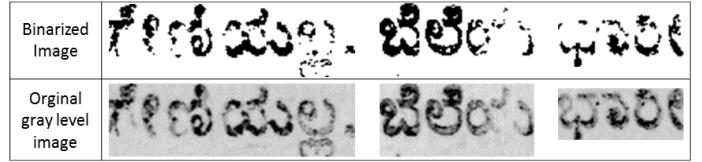


Fig. 2: Example word images that illustrate that stroke-connectivity information that was available in gray level image is sometimes lost during binarization.

II. ADAPTIVE NONLINEAR ENHANCEMENT OF THE IMAGE TO MINIMIZE CHARACTER SPLITS

Stroke-connectivity information that is there in the gray level document image is sometimes lost during binarization as shown by Fig. 2. Hence, we explored gray-level image enhancement techniques that can improve binarization to minimize character splits. Deepak and Ramakrishnan [26] proposed a technique to nonlinearly stretch the histogram of an image using power-law transformation (PLT) and applied it to obtain the best recognition result in the literature on the ICDAR2011 Robust Reading Competition Challenge-1: Word Recognition Task on born digital dataset. Later, they extended it to the different colour planes of camera-captured scene word images [27] and once again, obtained the best recognition results on the ICDAR2011 robust reading competition challenge 2. However, PLT was actually proposed with values of gamma greater than one, for the reverse problem of splitting character merges in born-digital word images. We propose a method for merging/handling character splits in degraded word images using PLT, where appropriate fractional values of gamma are chosen automatically by our algorithm for each word image, to obtain the best recognition result.

Figure 3 shows the flow chart of how gamma enhanced adaptive binarization is applied on each degraded image to obtain the best recognized result. The intensity of every pixel is modified as,

$$y_o[r][c] = (y_i[r][c])^\gamma \quad (1)$$

where $y_i[r][c]$ and $y_o[r][c]$ are the gray level intensities of the pixel at r^{th} row and c^{th} column, before and after gamma enhancement, respectively. Here, the intensities are expressed as a value in the interval $[0,1]$. The value of γ is varied in steps as decreasing integer powers of two, starting from $2^0 = 1$. The resulting gray level images are binarized using the Otsu global thresholding algorithm [28]. The recognized result with the best recognition score returned by our classifier is chosen as the final output. Figure 4 illustrates, with an example, the mechanism of how this improves the quality of a degraded Kannada word image. The gray level images obtained for different γ values, their histograms and the corresponding binarized images are all shown in Fig. 4.

Depending upon the level of degradation of the word images, different values of γ work optimally for different

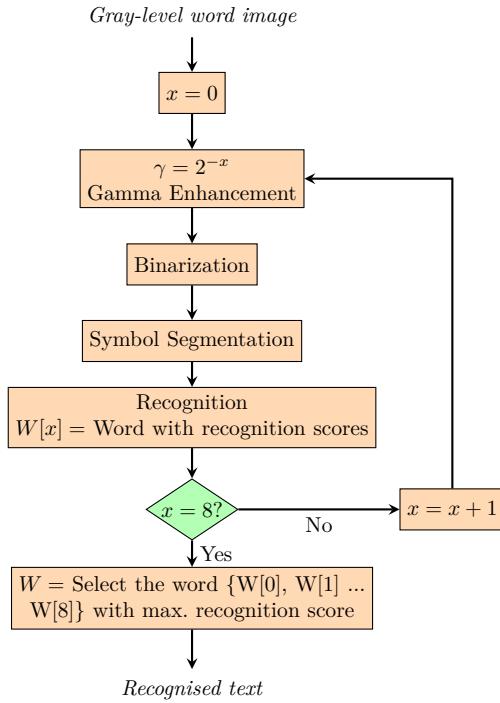


Fig. 3: Block schematic of Gamma-enhanced, locally adaptive binarization of gray level word images from scanned, old Kannada books, for improved OCR performance.

images. This is illustrated in Fig. 5, which shows seven word images with different levels of degradation, each of which is recognized correctly after enhancement with distinctly different values of γ .

III. NEW KANNADA DATASET FOR BENCHMARKING

For rigorously evaluating the performance of the proposed algorithm in handling character splits, we have created a benchmarking dataset of 1685 degraded word images obtained from old Kannada books. Figure 6 illustrates some sample images from this benchmarking dataset. The ground truth text for each of the test images is provided in a separate Unicode text file, as shown in Table I.

Performance of the proposed algorithm is evaluated using the Levenshtein distance between the recognized and the ground truth texts, across the entire benchmarking dataset. Let N denote the number of Unicodes in the ground-truth, and S , I and D denote the number of substitutions, insertions and deletions with respect to the ground truth, respectively. The Unicode recognition accuracy is determined as:

$$Accuracy = (N - S - I - D)/N \quad (2)$$

We are making this dataset publicly available at [6]. In Indian language processing, standardized datasets are a rarity. Hence, researchers report results on different datasets and it becomes difficult to assess the improvement

TABLE I: Filenames of some samples from the Kannada degraded word image dataset, together with the corresponding ground truth (GT) text for each test image. The GT has been provided in an accompanying, separate Unicode text file in the format shown below.

Image Name	Unicode Text (GT)	Image Name	Unicode Text (GT)
C0022.tif	ಇತಿಹಾಸದ	C0362.tif	ಅದಿಕ್ಷದಿಂದಾಗಿ
C0044.tif	ದೇಶಗಳ	C0365.tif	ಮುಂತಾದ
C0055.tif	ಗ್ರೀಕ್	C0386.tif	ಉಪಕಾರ
C0095.tif	ನುಗಳು	C0407.tif	ಮೂಲ
C0122.tif	ರಕ್ಷಣಾ	C0421.tif	ಮಟ್ಟ
C0250.tif	ಮೂರು	C0470.tif	ಸೇರಿದಂತे
C0274.tif	ಚೆಂಥಾವ್ಯಾ	C0508.tif	ಮತ್ತು

of the new algorithms they propose. Making our benchmarking dataset open is a step towards addressing that problem.

IV. RESULTS AND DISCUSSION

The Kannada word images in the created dataset have been recognized by a custom built SVM classifier, using wavelet and autocorrelation features, using the LIBSVM package [29]. Table II lists the recognition accuracies before and after the application of the adaptive enhancement technique on the images. The proposed technique is able to improve the performance of the classifier from 74.5% on the original images to 89.3% after our processing. There is a significant increase of 14.8% in the overall accuracy at the Unicode level, which is encouraging.

We have tracked the actual value of γ that resulted in the best recognition of each of the test images. Figure 7 shows the histogram of the different values of γ automatically selected by our algorithm in enhancing the 1685 degraded images in the test database. It is interesting to see that the entire range of possible values has been used. Each step change in the value of γ decreases the dynamic range of the histogram and after eight steps, as shown in Fig. 4, there is no more information left to enhance.

Since Tesseract OCR supports Kannada script, we ran Tesseract v3 and v4 on each of the 1685 degraded word images in the benchmarking dataset, and computed the character recognition accuracy as per eqn. (2). Also, after we obtained the best enhanced version of each word image (based on our algorithm), we again ran it through both the versions of Tesseract Kannada OCR. Table III shows the recognition accuracies of Tesseract v3 on the test dataset, before and after applying our nonlinear enhancement on the word images. Our gamma enhanced binarization entails an increase of 5.6% in the recognition accuracy of Tesseract v3. Though Tesseract v4.0.0 gives 75.7% accuracy before enhancement, the accuracy after enhancing the images is unexplainably low (40.7%), and hence, we have not reported it in Table III.

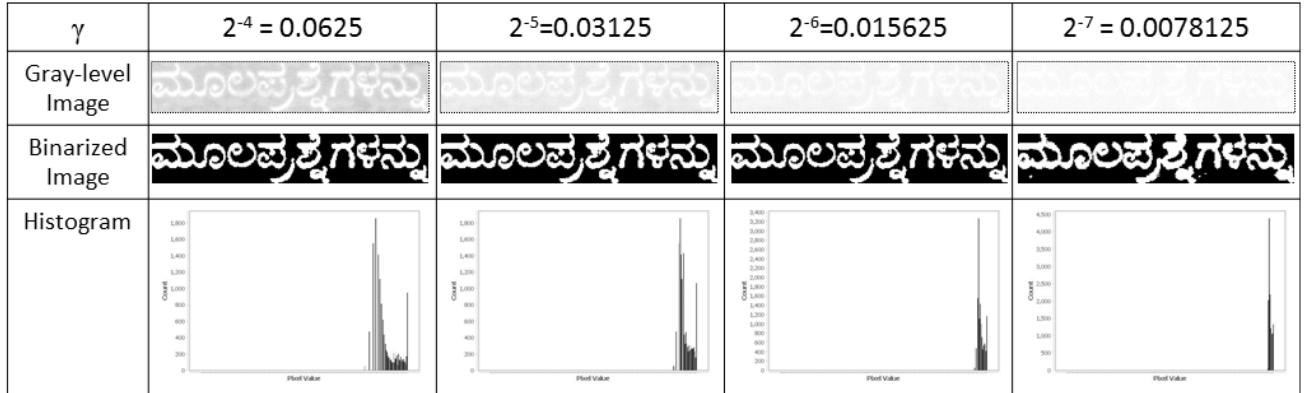
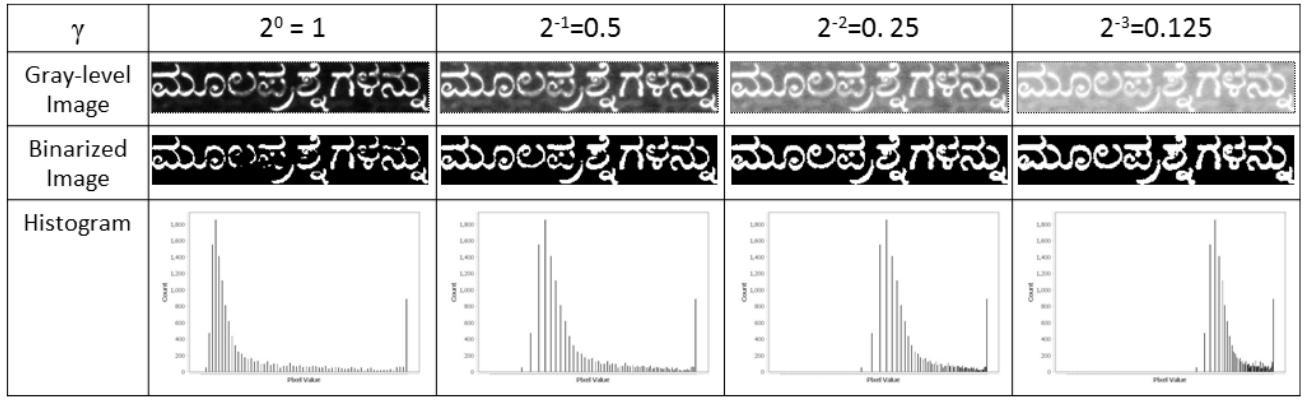


Fig. 4: Illustration of the impact of the nonlinear enhancement due to a range of gamma values on the gray-level image of a degraded Kannada word with splits, its histogram and the corresponding binarized output image.

γ	Gray-level degraded image	Otsu binarized word image	Binarized image after GEB
$2^{-1} = 0.5$			
$2^{-2} = 0.25$			
$2^{-3} = 0.125$			
$2^{-4} = 0.0625$			
$2^{-6} = 0.015625$			
$2^{-7} = 0.0078125$			
$2^{-8} = 0.0039062$			

Fig. 5: Illustration of the fact that different words with varying levels of degradation require different values of gamma for optimal binarization. The last column shows the best binarized image, as determined by the recognition score and the corresponding recognized word.

Gray-level Test Image	Otsu Binarized Image	Gray-level Image	Otsu Binarized

Fig. 6: Some samples from the benchmark dataset created by us, containing degraded (with cuts) word images of Kannada language. The corresponding binary images, obtained by Otsu global thresholding, are also shown, illustrating the level of splits in the different characters.

TABLE II: Comparison of the Unicode recognition accuracies (in %) before and after gamma enhanced binarization (GEB), on Kannada benchmarking dataset of 1685 degraded word images. N, M: # of Unicodes in ground-truth and OCR output. N = 15,486. S, I and D: # of Unicode substitutions, insertions and deletions.

	Accuracy	M	S	I	D
Before GEB	74.5	17,058	1,962	1,782	210
After GEB	89.3	15,954	927	599	131
Improvement	+14.8	-1,104	-1,035	-1,183	-79

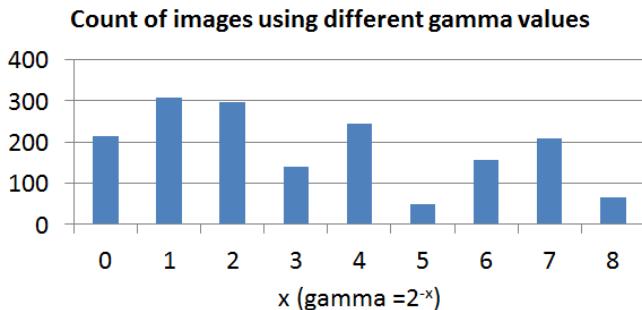


Fig. 7: Histogram of the different gamma values used by the algorithm for the images in the test database. It illustrates the extent to which different gamma values are automatically selected as being optimal for the different word images from the benchmark dataset of 1685 images.

TABLE III: Recognition accuracy (in %) of Tesseract OCR (v3.04.00) on the benchmark dataset, without and with gamma enhanced binarization (GEB). All the notations are the same as in Table II.

	Accuracy	M	S	I	D
Before GEB	36.7	19,136	5,582	3,937	287
After GEB	42.3	18,772	5,069	3,579	293
Improvement	+5.6	-364	-513	-358	+6

Figures 8 and 9 show the recognition results on all the representative samples from the benchmark Kannada degraded word image dataset, illustrated in Fig. 6, before and after the application of GEB. Errors in the recognized text are highlighted in red color. Figure 8 shows the cases, where GEB has managed to correct all the recognition errors that existed before applying our nonlinear enhancement to the word images. Figure 9 shows the cases, where the degradation/loss of information is so severe that even GEB does not suffice in the correct recognition of degraded words. In this figure, there are example images, where there is no improvement, there is partial correction of errors and also one case, where the error has increased after the automated, adaptive image enhancement.

Test Image – Otsu Binarized	-GEB Recognized Text	Best γ	GEB Image	+GEB Recognized Text
ದೇಶಗಳ	ದೇಶ/ನೂಳ	2^{-3}	ದೇಶಗಳ	ದೇಶಗಳ
ಸ್ಟೋ	ಸ್ಟೋ	2^{-8}	ಸ್ಟೋ	ಸ್ಟೋ
ನ:ಗಳು	ನಸೊಗಳು	2^{-1}	ನ:ಗಳು	ನಸೊಗಳು
ರಕ್ತೀಕಾರ	ರಕ್ತಿಂಬಿ	2^{-7}	ರಕ್ತೀಕಾರ	ರಕ್ತಿಂಬಿ
ಸೇರಿದಂತೆ:	ಸೇರಿದಂತೆ:	2^{-4}	ಸೇರಿದಂತೆ,	ಸೇರಿದಂತೆ,
ರೈಲ್	ರಲ್	2^0	ರೈಲ್	ರಲ್
ಭಾರೀ	ಭಾರೀ	2^{-6}	ಭಾರೀ	ಭಾರೀ
ಬೆಲೀಯು	ಬೆಲೀ	2^{-7}	ಬೆಲೀಯು	ಬೆಲೀಯು
ಗಿರಾಕಿಗಳು	,ರಜೆ(ಇ,	2^{-7}	ಗಿರಾಕಿಗಳು	ಗಿರಾಕಿಗಳು
ಕಾರಣ	ವಾತರಣ	2^{-4}	ಕಾರಣ	ಕಾರಣ
ಗೀರೀಳಿಯ್ಹ್ಯಾ	ರಾ/ಫಿಯ್ಹ್ಯಾ,,?	2^{-4}	ಗೀರೀಳಿಯ್ಹ್ಯಾ.	ಗೀರೀಳಿಯ್ಹ್ಯಾ.
ಕ್ಷಾಗಾರಿಕೆ	ಕ್ಷಾಲಾತರಿಕ	2^{-6}	ಕ್ಷಾಗಾರಿಕೆ	ಕ್ಷಾಗಾರಿಕೆ
ಹೆಡಿಕೆ	ಹೆಡಿಯೊ	2^{-7}	ಹೆಡಿಕೆ	ಹೆಡಿನ

Fig. 8: Illustration of some sample cases, where GEB has been successful in the correct recognition of the degraded words. - GEB: binarization of the original word image and the recognized word. + GEB: binarization after GEB and the word recognized. Errors in the recognized text in column 2 have been highlighted in red color.

Test Image – Otsu Binarized	-GEB Recognized Text	Best γ	GEB Image	+GEB Recognized Text
ಇತಿಹಾಸದೆ	ಇತಿಹಾಸದ	2^{-4}	ಇತಿಹಾಸದೆ	ಇತಿಹಾಸ
ನೇಮಾರು	ನೇಮಾರು	2^{-8}	ನೇಮಾರು	ನೇಮಾರು
ಬೆಂಥಾರ್	ಬೆಂಥಾರ್	2^{-1}	ಬೆಂಥಾರ್	ಬೆಂಥಾರ್
ಅದಿಕ್ಯಾದಿಂದಾಗಿ	ಅದಿಕ್ಯಾದಿಂದಾಗಿ	2^{-8}	ಅದಿಕ್ಯಾದಿಂದಾಗಿ	ಅದಿಗ್ರಿಂದಾಗಿ
ವಂತಾವೆ	ವಂತಾವೆ	2^{-2}	ವಂತಾವೆ	ವಂತಾವೆ
ಉಷಕಾರೆ.	ಉಷಕಾರೆ.	2^{-2}	ಉಷಕಾರೆ.	ಉಷಕಾರೆ.
ಮಾಲು	ಮಾಲು	2^{-2}	ಮಾಲು	ಮಾಲು
ಮುಖ್ಯ	ಮುಖ್ಯ	2^{-8}	ಮುಖ್ಯ	ಮಾಟ್
ವಂತು	ವಂತು	2^{-1}	ವಂತು	ವಂತು
ವಿಲ್ಲನೊ	ವಿಲ್ಲನೊ	2^{-4}	ವಿಲ್ಲನೊ	ವಿಲ್ಲನೊ
ವಿಕಾಳ	ವಿಕಾಳ	2^{-3}	ವಿಕಾಳ	ವಿಕಾಳ
ಇರ್ಲ	ಇರ್ಲ	2^{-1}	ಇರ್ಲ	ವಲರ್
ಒಂಟು	ಒಂಟು	2^{-5}	ಒಂಟು	ಸಿತು

Fig. 9: Illustration of some sample cases, where the degradation/loss of information is so severe that even after GEB, the word could not be correctly recognized. - GEB: binarization of the original word image and the word recognized. + GEB: binarization after GEB and the word recognized. Errors in the recognized text have been highlighted in red color. There are example images, where there is no improvement, there is partial correction of errors and also one case, where the error has increased after the automated, adaptive image enhancement.

V. CONCLUSION

Locally adaptive, nonlinear enhancement of gray-level word images has been proposed for improved binarization, and hence, better OCR recognition results on old Kannada documents, such as printed books. The proposed technique of gamma enhanced binarization has resulted in an improvement of 14.8% in the character recognition rate on the created benchmark dataset of 1685 degraded Kannada word images. Of course, there are severely degraded images, which could not be fully recognized correctly, even after the application of GEB. More work is needed to handle such difficult cases. The reported results do not make use of any dictionary or contextual information to correct the words, and the presence of the latter modules could improve the results further, when the proposed enhancement is embedded in a good OCR.

Even on the Google's Tesseract OCR for Kannada, our gamma enhanced binarization results in an improvement of 5.6% in the Unicode level accuracy.

Together with a good pre-processing that splits the merged characters [5], this enhancement is very promising to obtain better recognition performance from any Kannada OCR on old printed books of low quality. Further, the proposed technique is generic, and hence can be applied to degraded document images of any script, provided the character classifier used provides reliable recognition scores or posterior probabilities.

REFERENCES

- [1] N. Sinha and A. G. Ramakrishnan, "Blood cell segmentation using EM algorithm," in *Proc. Indian Conf. on Comp. Vision, Graphics and Image Processing (ICVGIP-02)*. ACM, 2002.
- [2] D. Kumar, M. N. Prasad, and A. G. Ramakrishnan, "MAPS: Midline analysis and propagation of segmentation," in *Indian Conf. Comp. Vision, Graphics and Image Proc.* ACM, 2012.
- [3] M. K. Jindal, R. K. Sharma, and G. S. Lehal, "A study of different kinds of degradation in printed Gurmukhi script," in *Computing: Theory and Applications. ICCTA'07. International Conf. on.* IEEE, 2007, pp. 538–544.
- [4] A. Madhavaraj, A. G. Ramakrishnan, H. R. Shiva Kumar, and N. Bhat, "Improved recognition of aged Kannada documents by effective segmentation of merged characters," in *Signal Processing and Communications (SPCOM), International Conference on.* IEEE, 2014, pp. 1–6.
- [5] H. R. Shiva Kumar, A. Madhavaraj, and A. G. Ramakrishnan, "Splitting merged characters of Kannada benchmark dataset using simplified paired-valleys and L-cut," in *Proc. 25th National Conference on Communication*, 2019.
- [6] MILE-IISc. (2018, Dec) Kannada benchmarking dataset of degraded word images, with character splits. [Online]. Available: <https://github.com/MILE-IISc/DegradedWordsKannada>
- [7] S. Dutta, N. Sankaran, K. P. Sankar, and C. V. Jawahar, "Robust recognition of degraded documents using character n-grams," in *Document Analysis Systems (DAS), 10th IAPR International Workshop on.* IEEE, 2012, pp. 130–134.
- [8] P. P. Kumar, C. Bhagvati, A. Negi, A. Agarwal, and B. L. Deekshatulu, "Towards improving the accuracy of Telugu OCR systems," in *Document Analysis and Recognition (ICDAR), Int. Conf. on.* IEEE, 2011, pp. 910–914.
- [9] D. Arya, C. V. Jawahar, C. Bhagvati, T. Patnaik, B. Chaudhuri, G. S. Lehal, S. Chaudhury, and A. G. Ramakrishnan, "Experiences of integration and performance testing of multilingual OCR for printed Indian scripts," in *Proc. Joint workshop on multilingual OCR and analytics for noisy unstructured text data.* ACM, 2011, p. 9.
- [10] A. G. Ramakrishnan and K. Mahata, "A complete OCR for printed Tamil text," in *Proc. Tamil Internet 2000.* INFITT, 2002, pp. 53–57.
- [11] K. G. Aparna and A. G. Ramakrishnan, "A complete Tamil optical character recognition system," in *International Workshop on Document Analysis Systems.* Springer, 2002, pp. 53–57.
- [12] A. Kokku and S. Chakravarthy, "A complete OCR system for Tamil magazine documents," in *Guide to OCR for Indic Scripts.* Springer, 2009, pp. 147–162.
- [13] B. Vijay Kumar and A. G. Ramakrishnan, "Machine recognition of printed Kannada text," in *International Workshop on Document Analysis Systems.* Springer, 2002, pp. 37–48.
- [14] ———, "Radial basis function and subspace approach for printed Kannada text recognition," in *International Conf. on Acoustics, Speech, and Signal Processing.* IEEE, 2004, pp. V–321.
- [15] T. V. Ashwin and P. S. Sastry, "A font and size-independent OCR system for printed Kannada documents using support vector machines," *Sadhana*, vol. 27, no. 1, pp. 35–58, 2002.
- [16] M. M. Prasad, M. Sukumar, and A. G. Ramakrishnan, "Divide and conquer technique in online handwritten Kannada character recognition," in *Proc. International Workshop on Multilingual OCR.* ACM, 2009, p. 11.
- [17] B. Nethravathi, C. Archana, K. Shashikiran, A. G. Ramakrishnan, and V. Kumar, "Creation of a huge annotated database for Tamil and Kannada OHR," in *Proc. Inter. Conf. on Frontiers in Handwriting Recognition (ICFHR).* IEEE, 2010, pp. 415–420.
- [18] D. Sachan, S. Dutta, T. Naveen, and C. Jawahar, "Segmentation of degraded Malayalam words: methods and evaluation," in *Computer Vision, Pattern Recog., Image Proc. and Graphics (NCVPRIPG), 3rd National Conf. on.* IEEE, 2011, pp. 70–73.
- [19] C. Sumetphong and S. Tangwongsan, "Recognizing broken characters in Thai historical documents," in *Advanced Computer Theory and Engineering (ICACTE), 3rd International Conf. on,* vol. 1. IEEE, 2010, pp. V1–99.
- [20] ———, "Effectively recognizing broken characters in historical documents," in *Computer Science and Automation Engineering (CSAE), IEEE Int. Conf. on,* vol. 3, 2012, pp. 104–108.
- [21] W. Peerawit, W. Yingsaeree, and A. Kawtrakul, "The utilization of closing algorithm and heuristic information for broken character segmentation," in *Cybernetics and Intelligent Systems, IEEE Conf. on,* vol. 2, 2004, pp. 775–779.
- [22] D. Yu and H. Yan, "Reconstruction of broken handwritten digits based on structural morphological features," *Pattern Recognition*, vol. 34, no. 2, pp. 235–254, 2001.
- [23] L. Likforman-Sulem and M. Sigelle, "Recognition of degraded characters using dynamic Bayesian networks," *Pattern Recognition*, vol. 41, pp. 3092–3103, 2008.
- [24] M. Droettboom, "Correcting broken characters in the recognition of historical printed documents," in *Digital Libraries, Joint Conf. on.* IEEE, 2003, pp. 364–366.
- [25] F. Drira, F. LeBourgeois, and H. Emptoz, "Document images restoration by a new tensor based diffusion process: Application to the recognition of old printed documents," in *Document Analysis and Recognition, ICDAR. 10th International Conf. on.* IEEE, 2009, pp. 321–325.
- [26] D. Kumar and A. G. Ramakrishnan, "Power-law transformation for enhanced recognition of born-digital word images," in *Signal Processing and Communications (SPCOM), International Conf. on.* IEEE, 2012, pp. 1–5.
- [27] D. Kumar, M. N. Anil Prasad, and A. G. Ramakrishnan, "NESP: Nonlinear enhancement and selection of plane for optimal segmentation and recognition of scene word images," in *Document Recognition and Retrieval XX*, vol. 8658. International Society for Optics and Photonics, 2013, pp. 865–806.
- [28] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [29] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines, software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>," 2001.

Full Reference Stereoscopic Video Quality Assessment Based On Spatio-Depth Saliency And Motion Strength

Sameeulla Khan, Md

Associate Professor

*Department of Electronics and Communication Engineering
Vellore Institute of Technology, Amaravati
email: sameeulla.k@vitap.ac.in*

Sumohana Channappayya

Associate Professor

*Department of Electrical Engineering
Indian Institute of Technology Hyderabad
email: sumohana@iith.ac.in*

Abstract—Stereoscopic video quality is a perceptual phenomena that is related to the human visual system (HVS). In this paper, we present a spatio-depth saliency and motion strength based full reference stereoscopic video quality metric (FRSVQA). Initially, we obtain a spatial distortion map on every video frame to estimate spatial quality. The spatial distortion map is then refined by the depth salient maps to estimate depth quality. We also estimate the temporal quality by refining the spatial distortion map with the inter-frame difference map at the locations specified by motion edges. The spatial, depth and temporal qualities are systematically combined and averaged over the frames to estimate the overall stereo video quality metric.

I. INTRODUCTION

The long history of stereoscopic/3D (S3D) image and video can be traced to the beginning of photography itself. In recent years, three-dimensional (3D) multimedia technologies have received wide attention as a result of a great impetus from the entertainment industry. Since 3D video is the combination of two single view video, its development and utility is on par with 2D multimedia technologies. Also the broadcasting of 3D content over the internet become common. With the increase of 3D capable phones and 3D broadcast services it is reasonable to believe that the consumption of 3D video will increases over the next few years. This requires the need of compression for storage and transmission of 3D video which degrades the perceptual quality. These advancements demands the objective quality assessment since subjective quality assessment on large volumes of data is time consuming and expensive. In this paper we propose an algorithm for full reference stereoscopic video quality metric (FRSVQA) which is adapted from the principles of saliency based stereoscopic image quality assessment. We briefly review the previous methods for stereoscopic video quality assessment algorithms to place our work in FRSVQA and allow for comparative analysis.

SVQA starts with the simple implementation of 2D objective VQA algorithms [2] [3] [4], which gives promising models for better FRSVQA algorithms. Later De Silva et al. [5] proposed an FRSVQA algorithm which is based on three perceptual features namely structural distortion, blur measurement and content complexity feature of reference and

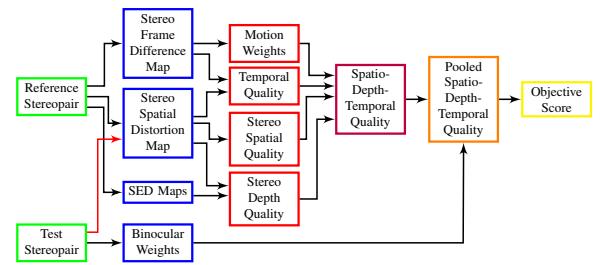


Fig. 1: Framework of the proposed FRSVQA algorithm.

test S3D video. Later these features are trained and aggregate to compute the quality score. Feng et al. [6] proposed a stereo just noticeable difference based (SJND) FRSVQA algorithm. It mainly includes temporal JND (TJND) and inter-view JND (IJND). In [7], Wang et al. proposed an FRSVQA algorithm. They predict the 3D quality scores from 2D video qualities of left and right view. Finally, the overall prediction of 3D video quality is calculated by a weighted average of the left- and right-view video quality. Appina et al. [8] proposed an FRSVQA based on the separable representation of motion and binocular disparity in the visual cortex. They used a structural based 2D IQA method for spatio-depth features and optical flow for temporal features. Hong et al. [9] proposed an objective metric to measure video compression distortion. They used stereo visual saliency based pooling approach to accumulate local spatial and temporal distortions. The proposed approach is novel in that it considers depth saliency and motion edges as a canonical factor in the FRSVQA task. We describe our approach in the following.

II. PROPOSED APPROACH

The proposed FRSVQA algorithm is based on stereo-depth saliency, that is successfully applied to FRSIQA [10]. With application to FRSIQA, [10] hypothesize that S3D image saliency can be separated and studied as two components: (i) image saliency and (ii) depth saliency. The SDSP saliency method [11] is used, with customized settings to predict the image saliency. The Canny image gradient is used for



(a) Frame of a video.



(b) Optical flow of the frame.

Fig. 2: Illustration of optical flow of a scene from [1]

structural comparison. Further, in [10], it is hypothesized that the notion of depth perception can be observed at a subset of edges of S3D image which is scientifically supported by [12], [13] and [14]. These edges are called salient edges with respect to depth perception (SED). The FRSIQA method in [10] is applied on each frame along with the addition of motion information. In this paper, we claim that to evaluate the temporal quality, it is enough to observe the variations of each test frame (with respect to reference frame) on motion edges.

Our algorithm involves estimating several components, per frame, that contribute to overall objective score. The sequence of obtaining each component is: (i) Spatial distortion map/Spatial quality, (ii) Depth quality, (iii) Temporal quality, (iii) Motion strength/weight and (iv) Spatio-depth-temporal quality. Fig. 1 shows the flow chart of our proposed approach. We explain our algorithm on the left reference and test video pair which is also applied on the right video pair. We consider both views during the pooling stage.

A. Spatial Distortion Map/Spatial Quality

Since a video is a stack of frames, a single frame can be interpreted as a 2D image. The spatial quality of each test frame can be obtained by using the principles of any conventional 2D FRIQA metrics. In this paper, we used the method described in [10] for estimating luminance quality of single view of a stereopair. As a result, we obtain the spatial distortion map of x^{th} test frame by comparing it with the corresponding x^{th} reference frame.

Let $Q_{Map_{lx}}$ be the spatial distortion map of the x^{th} frame of left test video with respect to its corresponding reference frame. Then the corresponding spatial quality is estimated as

$$S_{lx} = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N Q_{Map_{lx}}(i, j), \quad (1)$$

where $M \times N$ is the size of each frame.

B. Depth Quality

The estimation of depth quality is similar to [10] which uses SED maps on frame basis. Let SED_{lx} be the SED map of

x^{th} frame of pristine left video. Then the depth quality of x^{th} frame is estimated similar to [10], and is given by

$$S_{dlx} = \frac{\sum_{i=1}^M \sum_{j=1}^N SED_{lx}(i, j) Q_{Map_{lx}}}{\sum_{i=1}^M \sum_{j=1}^N SED_{lx}(i, j)}. \quad (2)$$

Similarly, we will also obtain S_{dlx} with its corresponding SED map as SED_{rx} .

C. Temporal Quality

As spatial distortion map compares all the regions between reference and test frames, the temporal quality can be observed across those regions where there is motion information. Optical flow is the successfully applied parameter for estimating temporal quality [15]. Fig. 2 shows the optical flow of a scene. From Fig. 2b, it is evident that optical flow enhances the dense motion of the objects in Fig. 2a. Since HVS is more sensitive to structural changes [16], to quantify temporal quality we mainly focus on the edge map of optical flow, called motion edges. Because while viewing live streaming videos, we have observed that the temporal distortion mostly affects the edges of moving objects than the edges of non-moving objects. Further, area V2 represents motion boundary information and an important cue for shape recognition of moving object [17]. This motivates us to consider the motion edges as the primitive for estimating temporal quality.

The edge map of optical flow may not locate the exact location of motion edges because of computational issues. Hence, we consider the edge map of inter-frame difference maps as motion edges. Prior to computing inter-frame difference, inspired from Canny gradient [18], all the frames are smoothed with Gaussian low pass filter (LPF). In Fig. 3, we present different frame positions of a left pristine video and their corresponding frame difference maps (for x^{th} frame its corresponding frame difference map is obtained between x^{th} frame and $(x-1)^{th}$ frame). As we can see, there is motion in Figs. 3a and 3c since the car and gate are moving and in Fig. 3b, the objects in the scene are almost static. So, we believe that inter-frame difference map is the appropriate representative for motion information.

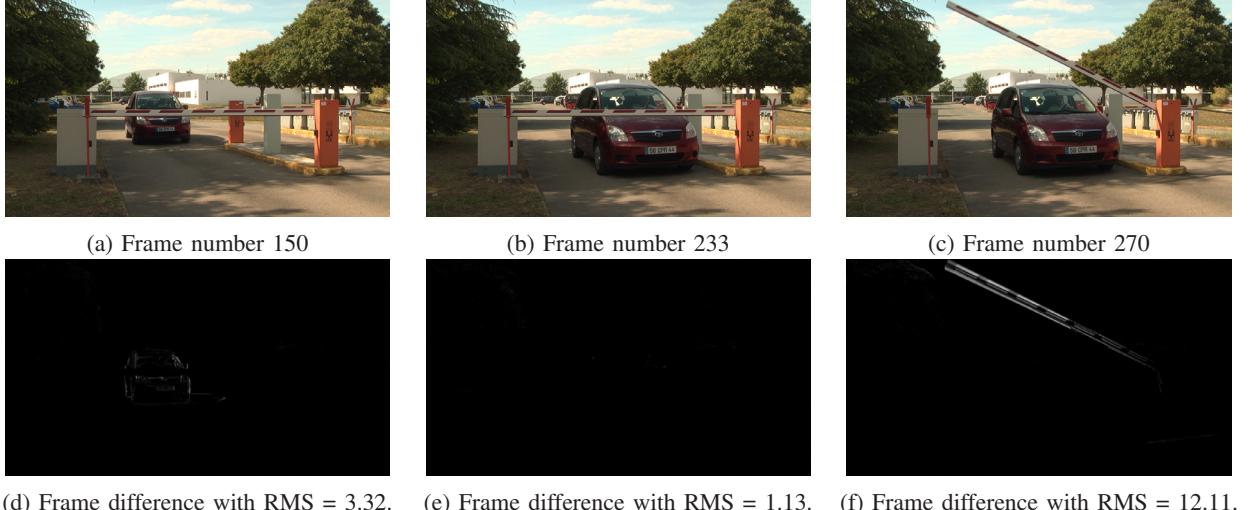


Fig. 3: Illustration of motion strength using frame difference.



Fig. 4: Motion Edge.

Let Δ_{lx} be the absolute frame difference map obtained between the x^{th} and $(x-1)^{th}$ smoothed frames of the left reference video. The value of $\Delta_{lx}(i, j)$ represents the strength of motion at the location (i, j) in the x^{th} frame. The edge map of Δ_{lx} represents the motion edges. We find the edge map of Δ_{lx} using the Canny edge detector. Fig. 4 shows the motion edge of the frame shown in Fig. 3a. The frame difference map, Δ_{lx} , is normalized with respect to the total sum at the motion edges, i.e.,

$$\delta_{lx} = \frac{\Delta_{lx} \times E_{D_{lx}}}{\sum_{i=1}^M \sum_{j=1}^N \Delta_{lx}(i, j) \times E_{D_{lx}}(i, j)}, \quad (3)$$

where $E_{D_{lx}}$ is the edge map of Δ_{lx} . The value of δ_{lx} at the motion edges represents the strength of motion. The temporal quality is then obtained by refining the spatial distortion map $Q_{Map_{lx}}$ with the weighted motion edges, δ_{lx} . i.e.,

$$S_{tlx} = \sum_{i=1}^M \sum_{j=1}^N Q_{Map_{lx}}(i, j) \times \delta_{lx}(i, j). \quad (4)$$

D. Motion Strength/Weight

We hypothesize that when there is no temporal variation, perceived quality is dominated by spatial quality. When there is a significant temporal variation, the observer's attention is drawn towards motion edges and is dominated by temporal

TABLE I: Performance evaluation of proposed approach over IRCCYN [19] database.

	PLCC	SROCC	RMSE
Q	0.9202	0.8918	0.4848
Q'	0.9061	0.8755	0.5241
Q_{ST}	0.8954	0.8896	0.5516
Q_{SD}	0.9186	0.8908	0.4895
Q_S	0.8943	0.8899	0.5542

quality. So, we customize the spatio-depth-temporal quality based on the motion strength. We consider the root mean square (RMS) value of Δ_{lx} , denoted by RMS_{lx} , as the representative for motion strength. In Fig. 3, the RMS values of the frame difference maps are listed below. The RMS value of Fig. 3d is more than Fig. 3f. This is because, in the former case the car's position is farther away from the camera view point and one would experience less motion for farther objects whereas in the latter case the gate is moving which is comparatively closer to the camera view point. Hence, RMS is a valid choice to represent the motion strength.

Using motion strength computed from RMS, the motion weight for the x^{th} frame is computed as

$$q_{lx} = \frac{RMS_{lx}}{255}. \quad (5)$$

This motion weight is used to prioritize either the temporal or spatio-depth quality.

E. Spatio-Depth-Temporal Score per Frame

In the previous sections, we compute spatial, depth and temporal qualities of the x^{th} frame of left test video with respect to its corresponding reference video. Now, we combine all these scores systematically. First, we compute spatio-depth score as

$$SD_{lx} = S_{lx} \times S_{dlx}. \quad (6)$$

Previously some authors fused the spatio-temporal features using convex combination for their applications ([22] [23]).

TABLE II: Comparison with state-of-the-art methods.

Algorithm	H264			JPEG2000			Overall		
	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE	PLCC	SROCC	RMSE
CHEN _{3D} [20]	0.7963	0.8035	2.5835	0.9358	0.8884	3.2863	0.8227	0.8201	2.9763
STRIQE _{3D} [21]	0.6836	0.6263	2.3683	0.8778	0.8513	3.2121	0.7599	0.7525	2.8374
FLOSIM _{3D} [8]	0.9589	0.9478	0.3863	0.9738	0.9548	0.2976	0.9178	0.9111	0.4918
3-D-PQI [9]	0.9306	0.9239	—	0.9413	0.9266	—	0.9009	0.8848	—
Proposed	0.9369	0.9201	0.3954	0.9706	0.9061	0.3154	0.9202	0.8918	0.4848

Inspired from them, we compute the spatio-depth-temporal score as

$$SDT_{lx} = (1 - q_{lx}) \times SD_{lx} + q_{lx} \times S_{tlx}. \quad (7)$$

Similarly we estimate SDT_{rx} on the x^{th} frame of the right test video.

F. Consolidated Scores And Final Score

For the x^{th} stereo frame, we have two spatio-depth-temporal scores. These scores are consolidated to have one score per frame. The average of the left and right scores, in the case of asymmetric distortions, will result in either over or under estimating the quality score. So, we follow the weighting strategy that was implemented in [10], where a geometric weighted combination is used. Therefore, the consolidated spatio-depth-temporal quality score is given by

$$SDT_x = SDT_{lx}^{e_{lx}} \times SDT_{rx}^{e_{rx}}, \quad (8)$$

where $e_{lx} + e_{rx} = 1$ and are derived in a similar manner as [10]. The overall quality of a test stereo video is computed as

$$Q = \frac{1}{P} \sum_{x=1}^P SDT_x, \quad (9)$$

where P is the number of frames.

Analogous to Q , we also compute four separate quality scores as following: (i) we compute the final score as Q' , by taking the average of spatio-depth and temporal quality instead of convex combination in Eq. 7, (ii) we compute the final score as Q_{ST} , by ignoring the depth quality in Eq. 6 (i.e., $SD_{lx} = S_{lx}$), (iii) we compute the final score as Q_{SD} , by ignoring the temporal quality in Eq. 7 (i.e., $SDT_{lx} = SD_{lx}$) and (iv) we compute the final score as Q_S , by ignoring both depth and temporal qualities in Eq. 7 (i.e., $SDT_{lx} = S_{lx}$).

III. RESULTS AND DISCUSSION

For the performance evaluation of the proposed algorithm, we used only the IRCCYN S3D VQA database [19], due to unavailability of other S3D VQA databases. It has five types of distortions namely H264, JPEG2000, image sharpening, reduction of resolution (RR) i.e., distortion introduced because of downsample-upsample and combo of the third and four. Following the trend in FRSVQA algorithms, we also report our results on H264 and JPEG2000 distortions. The performance analysis was carried out using standard statistical measures namely Pearson linear cross correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC) and root mean square error (RMSE). Higher values of PLCC and

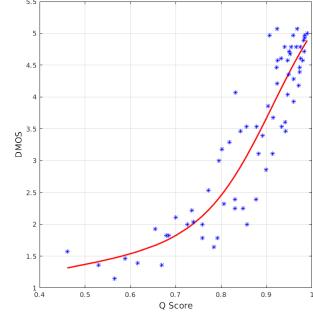


Fig. 5: Scatter plot over IRCCYN database.

SROCC and lower value of RMSE indicates better performance. We used the same subjective scores that are reported in [8]. All the results are reported post logistic fitting using five parameter recommended by Video Quality Experts Group [24].

Table I shows the performance of the the proposed method where we report the results for Q , Q' , Q_{ST} , Q_{SD} and Q_S . From the Table I, it is clearly evident that the systematic combination of spatio-depth-temporal quality (Q) result in the better performance over the other combinations. This reflects the importance of depth quality, motion weight and temporal quality. Fig. 5 shows the scatter plots of the proposed metric Q over IRCCYN database where the objective scores are almost linear with subjective scores. Table II shows the comparison of proposed algorithm with other methods over different distortions of IRCCYN database. The methods CHEN_{3D} and STRIQE_{3D} are the extensions of FRSIQA algorithms [21] and [20] respectively. [15] is an extension from 2D VQA to 3D VQA. All these algorithms ([21], [20] and [15]) are defined in [8]. From Table II, it shows that the proposed algorithm exhibit the state-of-the-art performance. Also our method is comparatively faster because of the following reasons: (i) We consider motion edges of the inter-frame difference as the temporal feature which is computationally inexpensive. (ii) We only used the reference disparity map values at the edge locations specified by SED maps, which greatly reduces the computational complexity of finding disparity maps between stereo frames of test stereoscopic video.

IV. CONCLUSION

In this paper, we present a full reference stereoscopic video quality assessment algorithm, which is systematically derived from spatial, depth and temporal qualities of each stereo frame. The saliency based spatial and depth quality is successfully

tested in full reference stereoscopic image quality assessment. We consider inter-frame difference of the reference stereo video as the motion constraint which is computationally inexpensive. Further, these frame difference maps are analyzed at the locations specified by the motion edges. Further, our spatial and depth qualities are also computationally less expensive. Our algorithm shows the state-of-the-art performance over IRCCYN S3D VQA database [19].

REFERENCES

- [1] "Motion Estimation and Optical Flow - Computer Vision Group, Freiburg," Available: <https://lmb.informatik.uni-freiburg.de/research/opticalflow/>.
- [2] C. Hewage, S. T. Worrall, S. Dogan, and A. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-d video," *Electronics letters*, vol. 44, no. 16, pp. 963–965, 2008.
- [3] P. Joveluro, H. Malekmohamadi, W. C. Fernando, and A. Kondoz, "Perceptual video quality metric for 3d video quality assessment," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*, pp. 1–4, IEEE, 2010.
- [4] C. D. M. Regis, J. V. de Miranda Cardoso, I. de Pontes Oliveira, and M. S. de Alencar, "Objective estimation of 3d video quality: A disparity-based weighting strategy," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2013 IEEE International Symposium on*, pp. 1–6, IEEE, 2013.
- [5] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an impairment metric for stereoscopic video: A full-reference video quality metric to assess compressed stereoscopic video," *IEEE transactions on image processing*, vol. 22, no. 9, pp. 3392–3404, 2013.
- [6] F. Qi, T. Jiang, X. Fan, S. Ma, and D. Zhao, "Stereoscopic video quality assessment based on stereo just-noticeable difference model," in *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pp. 34–38, IEEE, 2013.
- [7] J. Wang, S. Wang, and Z. Wang, "Asymmetrically compressed stereoscopic 3d videos: Quality assessment and rate-distortion performance evaluation," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1330–1343, 2017.
- [8] B. Appina, K. Manasa, and S. S. Channappayya, "A full reference stereoscopic video quality assessment metric," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 2012–2016, IEEE, 2017.
- [9] W. Hong and L. Yu, "A spatio-temporal perceptual quality index measuring compression distortions of three-dimensional video," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 214–218, 2018.
- [10] S. Khan and S. S. Channappayya, "Estimating depth-salient edges and its application to stereoscopic image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 12, pp. 5892–5903, 2018.
- [11] L. Zhang, Z. Gu, and H. Li, "Sdsp: A novel saliency detection method by combining simple priors," in *20th IEEE International Conference on Image Processing (ICIP), 2013*, pp. 171–175, IEEE, 2013.
- [12] B. Gillam, T. Flagg, and D. Finlay, "Evidence for disparity change as the primary stimulus for stereoscopic processing," *Perception & Psychophysics*, vol. 36, no. 6, pp. 559–564, 1984.
- [13] B. Gillam and E. Borsting, "The role of monocular regions in stereoscopic displays," *Perception*, vol. 17, no. 5, pp. 603–608, 1988.
- [14] R. von der Heydt, H. Zhou, and H. S. Friedman, "Representation of stereoscopic edges in monkey visual cortex," *Vision research*, vol. 40, no. 15, pp. 1955–1967, 2000.
- [15] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, 2016.
- [16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [17] M. Chen, P. Li, S. Zhu, C. Han, H. Xu, Y. Fang, J. Hu, A. W. Roe, and H. D. Lu, "An orientation map for motion boundaries in macaque v2," *Cerebral Cortex*, vol. 26, no. 1, pp. 279–287, 2014.
- [18] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 679–698, 1986.
- [19] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutierrez, and N. Garcia, "Nama3ds1-cospad1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3d stereoscopic sequences," in *2012 Fourth IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 109–114, 2012.
- [20] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.
- [21] S. K. Md, B. Appina, and S. S. Channappayya, "Full-reference stereo image quality assessment using natural stereo scene statistics," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 1985–1989, 2015.
- [22] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, 2011.
- [23] B. Wu, L. Xu, L. Zeng, Z. Wang, and Y. Wang, "A unified framework for spatiotemporal salient region detection," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 16, 2013.
- [24] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii (fr_tv2)," ftp://ftp.its.blrdrc.gov/dist/ituvividq/Boulder_VQEG_jan_04/VQEG_PhaseII_FRTV_Final_Report_SG9060E.doc, 2003, 2003.

Interpolated Compressed Sensing for Calibrationless Parallel MRI Reconstruction

Sumit Datta and Bhabesh Deka, SM, IEEE

Department of Electronics and Communication Engineering

Tezpur University, Tezpur-784028, Assam, India

E-mail: bdeka@tezu.ernet.in.

Abstract—Parallel magnetic resonance imaging (pMRI) in clinical study are commonly acquired in multiple slices; parallelly along different channels. Since, MRI traditionally suffers from slow data acquisition, reconstruction of images in clinical pMRI would be further slower. Compressed sensing MRI (CS-MRI) has successfully demonstrated its potential in reducing the scan time of pMRI by manifolds. Due to high correlation of adjacent slices in multislice sequence, interpolation of multi-slice data may be carried out to support non-uniform undersampling based CS reconstruction of slices in k-space. Exploiting intra/inter slice as well as multichannel data redundancy of multi-slice pMRI, it is possible to accelerate the scan time further. These correlations can be well modeled by introducing multidimensional wavelet forest sparsity and joint total variation regularization during the CS reconstruction. To validate our claim, a number of experiments are carried out with real pMRI datasets and results are compared with the state-of-the-art.

I. INTRODUCTION

Magnetic Resonance imaging (MRI) is one of the most preferred medical imaging modalities, especially for soft tissues like brain, abdomen, etc. due to its ability to provide good contrast without using ionizing radiations. But, it is slow for its low acquisition speed compared to its closest competitor i.e. computed tomography (CT). Lustig *et al.* introduced compressed sensing MRI (CS-MRI) [1] which makes reduction of MRI scan time possible without compromising its quality by measuring partial Fourier data and then applying some highly nonlinear algorithms for reconstruction of images.

In clinical practice, for better analysis of human organ 3D MRI is most commonly used. In most cases, a number of 2D slices are acquired from a small volume with no or very negligible inter-slice gaps maintaining the slice thickness around 2-3 mm. To apply CS for 3D MRI reconstruction, these 2D multislice data are to be undersampled in the frequency domain (k-space). It has been found that recent works utilize inter-slice correlation of 2D multi-slice data to estimate missing k-space samples. Pang and Zhang [2], Pang *et al.* [3], and very recently, Datta and Deka [4], [5] used the concept of interpolated compressed sensing (iCS). In iCS MRI, some slices are highly undersampled compared to others. Then, missing samples of a highly undersampled slice (H-slice) are estimated from samples of neighboring lightly undersampled slice (s) (L-slice (s)). It significantly reduces the

overall undersampling ratio, which means less scan time, better patient comfort and effective utilization of the MRI facility.

In parallel MRI (pMRI), to speed up the scan time multiple receiver coils (or channels) are used. Each receiver coil acquires only a fraction of whole k-space data. The target MR image is reconstructed using information of all receiver coils. There are broadly two types of approaches depending on the requirements of coil sensitivity information. Some methods explicitly require sensitivity information while others need the information implicitly. The main disadvantage of the former is that, in practice, it is quite difficult to estimate the coil sensitivity with good accuracy. A small error in the sensitivity profile may lead to significant artifacts in the reconstructed image. Methods, like, sensitivity encoding (SENSE) [6] and simultaneous acquisition of spatial harmonics (SMASH) [7] require sensitivity information explicitly. SENSE iteratively solves an inverse problem to obtain an MR image from undersampled k-space data of all receiver coils with an assumption that individual coil sensitivity information is available. Mathematically, it gives optimal reconstruction results if the coil sensitivity information is accurate. On the other hand, methods like generalized autocalibrating partially parallel acquisition (GRAPPA) [8] and AUTO-SMASH [9] implicitly use coil sensitivity information from acquired data, i.e. this methods are autocalibrating. In GRAPPA, first full k-space samples corresponding to a coil are estimated from undersampled k-space data of the remaining coils in the ensemble.

After development of CS-MRI, two CS-based pMRI approaches, namely, iterative self-consistent parallel imaging reconstruction from arbitrary k-space (SPIRiT) [10], CS-SENSE [11], and ESPIRiT [12] are invented which are basically autocalibrating methods. These methods respectively estimate interpolation weights and sensitivity maps from the full acquired center region of the k-space. Recently, some calibration-less work [13]–[15] are also reported, meaning they do not require any calibration information explicitly or implicitly.

A standard CS MRI reconstruction problem consists of wavelet domain sparsity and spatial domain gradient sparsity terms along with data fidelity term [1]. Chen and Huang [16] demonstrated that MR images are not only sparse in the wavelet domain but also follow quadtree structure, i.e. if any parent coefficient in the coarse scale is large (or small) then corresponding four children coefficients in the adjacent finer scale are also large (or small). This is also

known as wavelet domain quadtree structure. In case of multi-channel MRI, corresponding images of adjacent channels are highly correlated, thus wavelet coefficients of adjacent channel images of identical positions are expected to be similar. Chen *et al.* [17] utilize wavelet domain inter-channel similarity to explore sparsity during multi-channel MR image reconstruction. Similarly, authors [4] also proposed a tree sparsity based approach for CS based multi-slice MRI reconstruction.

In this paper, we have proposed a multi-slice parallel MRI reconstruction method using the CS. We formulate the above problem using a 4D CS MRI reconstruction model that is capable of exploiting data redundancy both in intra/inter-slice (3D multi-slice data) and inter-channel simultaneously (4D multi-slice parallel MRI data). In addition, to take advantage of gradient domain sparsity of MRI images, a 4D joint total-variation (TV) regularization is imposed on the 4D data. In order to exploit the advantage of k-space interpolation as in [3]–[5], we also apply interpolation to estimate missing k-samples of under-sampled multi-slice data. This step significantly enhances the speed of multi-slice data acquisition. For interpolation, we have adopted the method proposed in [4] for multi-slice data.

For comparison with existing CS based pMRI methods extensive simulations are carried out with two real 4D MRI datasets. Results are compared with some of the existing similar works in terms of both qualitative and quantitative evaluation metrics. We observe significant improvements in case of the proposed method over other compared methods.

The rest of the paper is organized as follows: Section I, discusses the related background. Section III describes the proposed technique, followed by simulation results and discussions in Section IV. Finally, Section V draws some conclusions of the proposed work and future research directions.

II. BACKGROUND

A. Compressed Sensing parallel MRI

In CS based pMRI reconstruction a standard minimization problem may be defined as follows- suppose $\mathbf{y}_c \in \mathbb{R}^m$ is the undersampled k-space data of aliased image of the c^{th} coil, $\mathbf{x}_c \in \mathbb{R}^n$ i.e. $\mathbf{y}_c = \mathbf{F}_u \mathbf{x}_c$, where $m << n$, $c = 1, 2, \dots, C$; C the total number of coils, and $\mathbf{F}_u \in \mathbb{R}^{m \times n}$ is a partial Fourier operator [11]. Now, to reconstruct \mathbf{x}_c from given \mathbf{y}_c , one need to solve the following optimization problem:

$$\hat{\mathbf{x}}_c = \underset{\mathbf{x}_c}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{F}_u \mathbf{x}_c - \mathbf{y}_c\|_2^2 + \lambda_1 \|\Psi \mathbf{x}_c\|_1 + \lambda_2 \|\mathbf{x}_c\|_{\text{TV}}, \quad (1)$$

where λ_1 and λ_2 are regularization parameters, Ψ denotes transform operator, like, wavelet, and $\|\mathbf{x}_c\|_{\text{TV}}$ represents the TV norm of \mathbf{x}_c - $\|\mathbf{x}_c\|_{\text{TV}} = \sum_{i,j} \sqrt{\{(\nabla_h \mathbf{x}_c)_{i,j}\}^2 + \{(\nabla_v \mathbf{x}_c)_{i,j}\}^2}$ with ∇_h and ∇_v the first-order finite difference operators in two axial directions, respectively.

Most of the existing pMRI methods require coil sensitivity information either explicitly or implicitly. Reconstruction quality of these methods directly depends on the calibration

data and approach of estimation. Majumdar and Ward [13] proposed a calibration less multi-coil CS MRI reconstruction method, namely, the *calibration-Less Multi-coil MRI* (CaLM-MRI). This approach does not require estimation of coil sensitivity information. Here, first individual coil images are reconstructed using CS MRI reconstruction corresponding to all receiver coils and then these images are combined using sum-of-square to get the final image. Similar calibration less approach are reported in [15], [17]. They experimentally demonstrated that joint TV-norm regularization based joint CS MRI reconstruction gives better results compared to other well known pMRI methods, namely, GRAPA, SPIRiT, CS-SENSE, and CaLM-MRI.

Chen *et al.* [17] introduce the forest sparsity for pMRI. In pMRI, channel images are highly correlated because all of them represent the same field of view (FoV) with different spatial sensitivities. Therefore, location of edges and boundaries are in the same positions with different pixel intensities. Hence, wavelet coefficients of different channel images have a similar pattern. Chen *et al.* utilize this prior knowledge of data redundancy in multiple channel images to improve quality of CS-MRI reconstruction. Similarly, authors in [4] used forest sparsity concept for exploiting data redundancy in adjacent slice images in multi-slice 3D MRI.

B. Interpolation in multi-slice MRI

Recently, in [4] authors proposed a k-space domain 3D interpolation scheme for multi-slice MRI. In this method, first, a virtual L-slices (VL-slice) is estimated from two neighboring L-slices and kept it in between two groups of H-L-H. Finally, samples of any target H-slice are estimated using samples of adjacent L and VL slices. This method gives better quality of interpolation due to accurate bidirectional estimation of k-space samples. Further, it is computationally faster compared to the state-of-the-art.

III. PROPOSED METHOD

Although, wavelet domain forest sparsity and spatial domain joint TV-norm sparsity are previously used but none of them considered extending these concepts to exploit data redundancy existing in multi-slice parallel MRI data. However, MRI scanners in clinical practice (e.g. 1.5T GE Signa HDxt, 3.0T GE Discovery MR750) provide four dimensional multi-slice pMRI data. In this paper, we propose an interpolated CS based calibration less multi-slice pMRI reconstruction method. It consists of two stages: Interpolation and CS Reconstruction. For k-space interpolation, we extend the technique proposed in [4] for multi-slice pMRI. On the other hand, for CS reconstruction after interpolation, we propose a novel four dimensional wavelet domain forest sparsity and joint TV-norm based CS-MRI reconstruction technique.

A. Multi-slice pMRI k-space Interpolation

After transforming the multi-slice MRI sequence into k-space, we apply non-uniform undersampling across different slices in such a way that in every three adjacent slices first one

is L-slice and remaining two are H-slices. For example, suppose that there are 10 slices then undersampled slice sequence is (L-H-H)-(L-H-H)-(L-H-H)-L. The same slice-sequence is to be maintained for different channels. Since, the line type of undersampling strategy is the simplest and easily implemented in hardware, we consider it for k-space undersampling. As reported in [4], for the L-slice, a few consecutive rows from the center along with some randomly selected rows from the periphery of the k-space are acquired. On the other hand, in case of the H-slice, a few consecutive rows are acquired from the center region only. Now, we estimate a VL-slice from samples of two neighboring L-slices by weighted averaging approach detailed in our previous work [4]. After estimation of VL-slice, we kept it in between two H-slices and interpolate samples of the H-slices using the nearest L-slice and VL-slice. The same process is followed for other H-slices as well. After interpolation, all L and interpolated H-slices in the multi-slice sequence contain the same amount of undersampled data on which CS reconstruction is to be applied.

B. Joint CS pMRI Reconstruction

Undersampled 4D k-space data denoted by $\mathbf{Y} = \{(\mathbf{y}_{1,1}^L, \mathbf{y}_{1,2}^H, \dots, \mathbf{y}_{1,C}^L); (\mathbf{y}_{2,1}^L, \mathbf{y}_{2,2}^H, \dots, \mathbf{y}_{2,C}^L); \dots; (\mathbf{y}_{S,1}^L, \mathbf{y}_{S,2}^H, \dots, \mathbf{y}_{S,C}^L)\}$, is being obtained by $\mathbf{Y} = \mathbf{F}_u \mathbf{X}$ where C and S denotes total number of channels and slices, respectively. The corresponding image domain data are: $\mathbf{X} = \{(\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \dots, \mathbf{x}_{1,C}); (\mathbf{x}_{2,1}, \mathbf{x}_{2,2}, \dots, \mathbf{x}_{2,C}); \dots; (\mathbf{x}_{S,1}, \mathbf{x}_{S,2}, \dots, \mathbf{x}_{S,C})\}$.

To exploit joint intra/inter-slice and inter-channel data redundancy, we propose a novel joint 4D CS reconstruction model by extending the single-vector model in Eq. 1 to a multi-vector model capable of exploiting joint correlations among different slices/channels besides intra slice/channel correlations of pMRI data. It consist of two regularization terms; one controlling the wavelet domain forest sparsity and the other controlling the gradient sparsity using the joint TV-norm. So, we write:

$$\hat{\mathbf{X}} = \arg \min_{\mathbf{X}} \frac{1}{2} \sum_{s=1}^S \sum_{c=1}^C \|\mathbf{F}_u \mathbf{x}_{s,c} - \mathbf{y}_{s,c}\|_2^2 + \lambda_1 \|\mathbf{X}\|_{\text{JTV}} + \lambda_2 \sum_{g_i \in G} \|(\Psi \mathbf{X})_{g_i}\|_2, \quad (2)$$

where first term is the data fidelity, λ_1 and λ_2 are regularization parameters, first regularization term denotes the wavelet domain forest sparsity in 4D computed by applying the $\ell_{1,2}$ -norm on wavelet coefficients of multi-slice pMRI; g_i is the i^{th} forest group and G is the set of all such groups, i.e. $G = \{g_1, \dots, g_i, \dots, g_p\}$. A particular 4D forest group contains parent-child pairs from multiple slices within multiple channels. This particular grouping arrangement exploits similarity among wavelet coefficients of same location from multiple slices of different channels during iterative group shrinkage operations. Similarly, the second regularization term is the joint TV-norm in 4D and defined as-

$$\|\mathbf{X}\|_{\text{JTV}} = \sqrt[n]{\sum_{i=1, j=1}^{S, C} \left\{ (\nabla_h \mathbf{x}_{s,c})_{i,j}^2 + (\nabla_v \mathbf{x}_{s,c})_{i,j}^2 \right\}}.$$

In multi-slice pMRI gradient of images are not only sparse but also jointly sparse along slices as well as channel directions.

Algorithm 1 Proposed Algorithm

Input: $\Psi, \mathbf{F}_u, \mathbf{Y}, \lambda_1, \lambda_2, \beta, L$

Initialization: $\{\mathbf{X}^0, \mathbf{r}^1\} \leftarrow \mathbf{F}_u^T \mathbf{Y}, \{t^1, k\} \leftarrow 1$

```

1: while until converge do
2:    $\mathbf{Z}^k \leftarrow \text{shrinkgroup} \left( G\Psi \mathbf{X}^{k-1}, \frac{\lambda_2}{\beta} \right)$ 
3:    $\mathbf{P}^k \leftarrow \mathbf{r}^k - \frac{1}{L} \left[ \sum_{s=1, c=1}^{S, C} \mathbf{F}_u^T (\mathbf{F}_u \mathbf{r}_{s,c}^k - \mathbf{Y}_{s,c}) + \beta \Psi^T G^T \{G\Psi \mathbf{r} - \mathbf{Z}^k\} \right]$ 
4:    $\mathbf{X}^k \leftarrow \arg \min_{\mathbf{X}} \frac{L}{2} \|\mathbf{X}^{k-1} - \mathbf{P}^k\|_2^2 + \lambda_1 \|\mathbf{X}^{k-1}\|_{\text{JTV}}$ 
5:    $\mathbf{X}^k \leftarrow \text{project} (\mathbf{X}^k, [l, u])$ 
6:    $t^{k+1} \leftarrow \left( 1 + \sqrt{1 + 4(t^k)^2} \right) / 2$ 
7:    $\mathbf{r}^{k+1} \leftarrow \mathbf{X}^k + \frac{t^k - 1}{t^{k+1}} (\mathbf{X}^k - \mathbf{X}^{k-1})$ 
8:    $k \leftarrow k + 1$ 
9: end while

```

Output: $\hat{\mathbf{X}} \leftarrow \mathbf{X}^k$

The above minimization problem is a composite regularization problem for CS reconstruction of pMRI without the coil sensitivity map in the data fidelity term. First, we decompose it into two subproblems, which can be efficiently solved using existing methods [18]. We summarize the algorithmic steps of the proposed technique in Algorithm-1. In the algorithm, the step 2 is corresponding to the wavelet domain 4D forest sparsity subproblem, denoted by a group shrinkage and thresholding operator “`shrinkgroup ()`” similar to the one described in [19]. On the other hand, the step 4 is corresponding to the joint TV subproblem. The `project ()` in step 5 is a function to scale the image intensity range in $[l, u]$. Here, G is a binary matrix, non-zero locations at each row of G represent locations of wavelet coefficients within a group. One of the key features of the proposed algorithm is the fast convergence due to the acceleration scheme as reported by [18], [20].

In clinical practice, a typical multi-slice pMRI data contains 50-300 slices. For simulation of such problem, we consider a number of smaller 4D subproblems, where each subproblem consists of multiple coil images of three adjacent slices. More details about the simulation is given in the next section.

IV. EXPERIMENTAL RESULTS

All simulation works of this paper are carried out in MATLAB environment on a Dell workstation equipped with Intel Xeon CPU E5-2650, 2.20GHz, 128GB of RAM. For simulation, we have collected two real MRI datasets, one from a local hospital and another from an online publicly available database. Dataset-I ($256 \times 256 \times 8 \times 130$), brain MRI collected from a local hospital. This is acquired with following settings: scanner: 1.5T GE Signa HDxt, slice thickness 0.6mm, TE: 2.268ms, TR: 5.736ms, flip angle: 65, receiver coil: 8HRBrain,

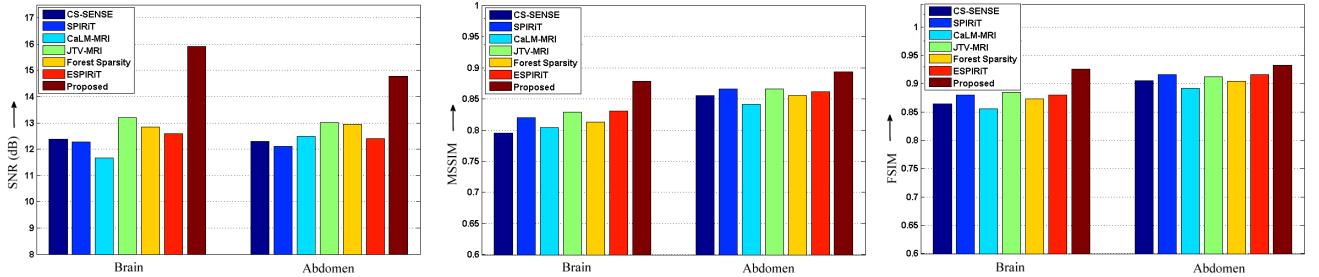


Fig. 1: Comparison of PSNR (dB), MSSIM, and FSIM values of reconstruction with 20% sampling ratio.

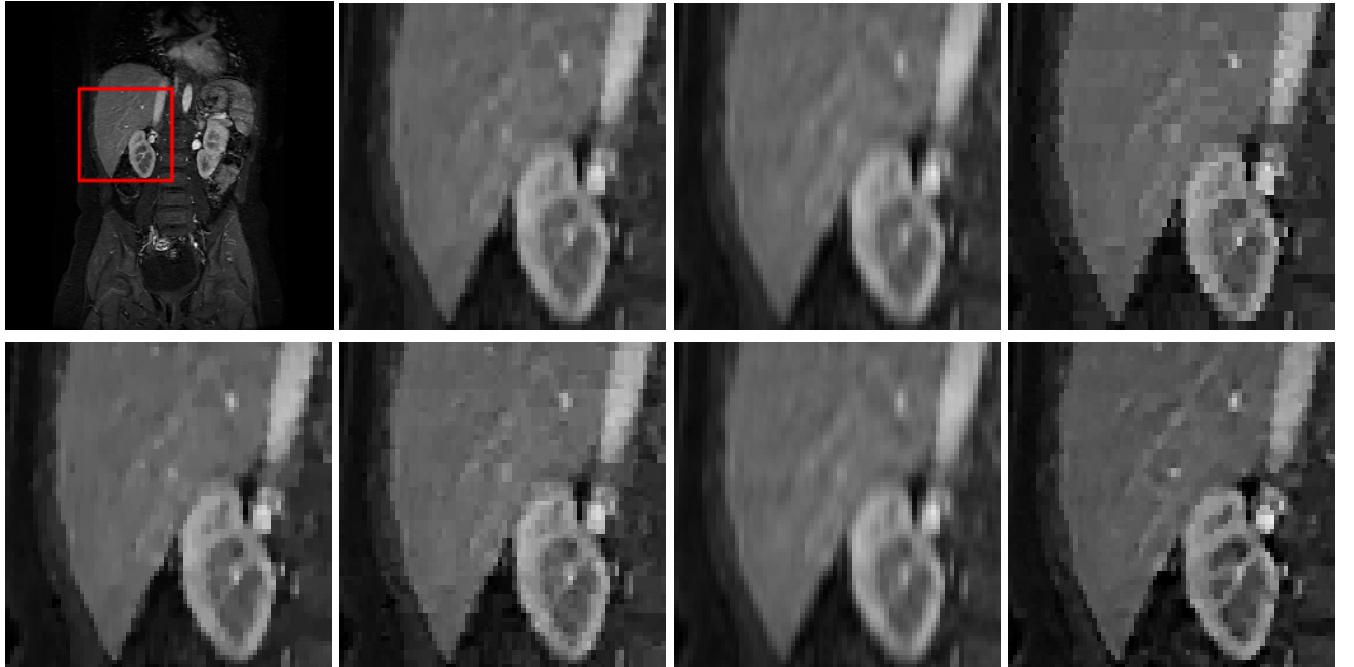


Fig. 2: Comparison of reconstructed Abdomen images using different techniques from 20% k-space samples. First row left to right: Original Abdomen image [using the full k-space] (slice #30) and reconstructed images by the CS-SENSE, the SPIRiT, and the CaLM-MRI respectively. Second row left to right: reconstructed images by the JTV-MRI, the Forest sparsity, the ESPIRiT and the Proposed technique, respectively

TABLE I: Comparison of computational cost in terms of CPU Time (in Hrs.)

Method	Dataset I ($256 \times 256 \times 8 \times 130$)	Dataset II ($256 \times 256 \times 8 \times 100$)
CS-SENSE [11]	8.86	7.13
SPIRiT [10]	1.67	1.22
CaLM-MRI [13]	3.56	2.68
JTV-MRI [15]	0.70	0.51
Forest Sparsity [17]	1.23	0.94
ESPIRiT [12]	1.55	1.17
Proposed	0.81	0.57

acquisition matrix: 256×256 . Dataset-II ($256 \times 256 \times 8 \times 100$) abdomen MRI collected online from ¹.

¹<http://old.mridata.org/undersampled/abdomens/>

For performance comparisons, we use only 20% k-space data for all methods. As mentioned earlier, in case of the proposed method undersampling is performed non-uniformly across slices in groups with each group containing three adjacent slices, i.e. L-H-H—L-H-H—... In case of an L-slice undersampling ratio is 30% and in case of an H-slice undersampling ratio is 15%, so the average undersampling ratio is only 20%. However, experimental are also carried out for other undersampling ratios as well. In this paper, we report results for average undersampling ratio of 20% only. We use “db2” mother wavelet with 4 levels decomposition in our simulations. Values of λ_1 and λ_2 are selected based on experimental results for efficient CS based MR image reconstruction reported in [18]. The results presented in this paper are average of 20 run in MATLAB. For comparison of reconstruction quality, three quality evaluation metrics,

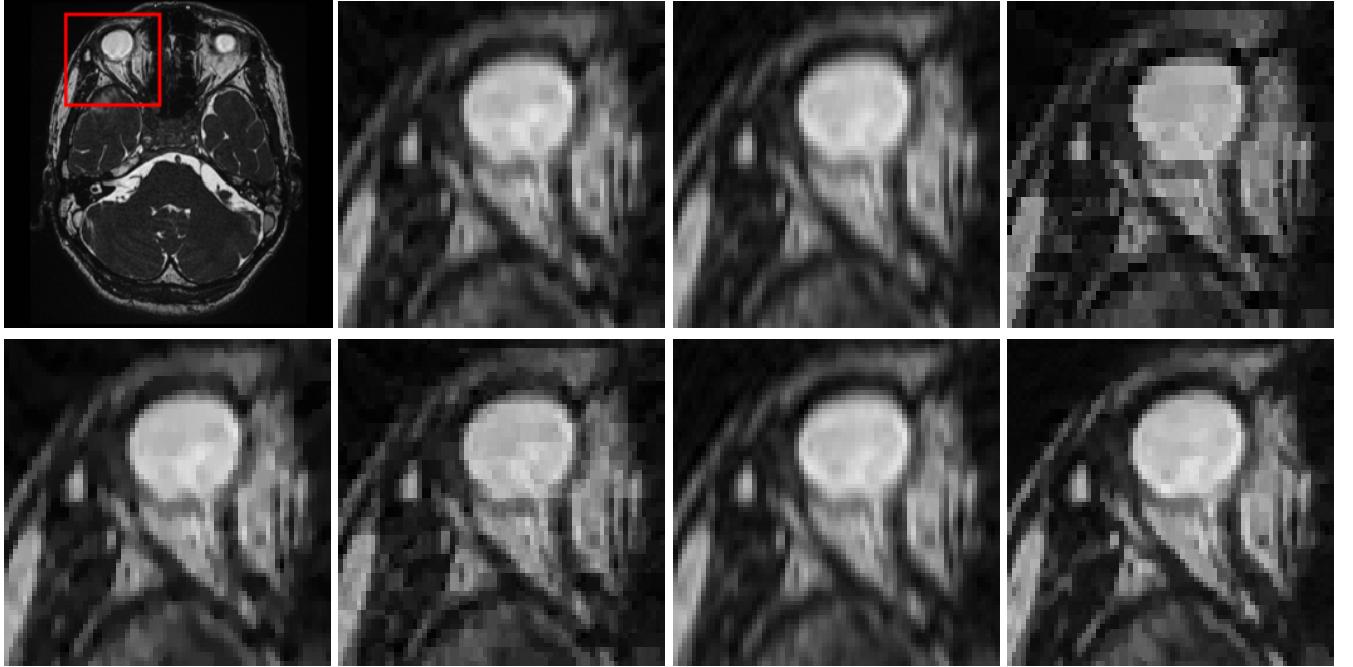


Fig. 3: Comparison of reconstructed Brain images using different techniques from 20% k-space samples. First row left to right: Original Brain image [using the full k-space] (slice #64) and reconstructed images by the CS-SENSE, the SPIRiT, and the CaLM-MRI respectively. Second row left to right: reconstructed images by the JTV-MRI, the Forest sparsity, the ESPIRiT and the Proposed technique, respectively

namely, signal-to-noise ratio (SNR), mean structural similarity index (MSSIM), and feature similarity index (FSIM) [21] are used. Further, computational cost of different techniques are measured in terms of the CPU time.

Comparison of CS-MRI reconstruction quality in terms of SNR, MSSIM, and FSIM are shown in Fig. 1. From the figure we observe that the proposed method gives an improvement of SNR 2.5dB and 2dB respectively for brain and abdomen datasets. Similar significant improvements also observe in terms of MSSIM and FSIM for both datasets. For better visualization a cropped portion of a randomly selected slice from both datasets for all compared methods are shown in Figs. 2-3. From the figure it is clearly visible that the proposed method gives better reconstruction in terms of preservation of edges with less amount of aliasing artifacts compared to other methods.

Computational cost in terms of CPU Time is shown in Table I. From the table we observe that CS-SENSE takes longer computation time among all compared method. On the other hand, the JTV-MRI and the proposed method takes less computation time compare to other methods. Although, the proposed method takes slightly higher computation time than the JTV-MRI but reconstruction quality is significantly better than that of the JTV-MRI.

V. CONCLUSION

We have proposed a novel CS based multi-slice pMRI reconstruction method for clinical MRI datasets. The proposed

reconstruction model is equipped to handle both intra/inter slice and inter channel data redundancy simultaneously. To validate the proposed model and to demonstrate its efficacy results are compared with a few other similar existing methods for two datasets. Further works are being carried out to enhance the acceleration of the proposed method by hardware implementation using GPGPU.

VI. ACKNOWLEDGMENT

Authors would like to thank for the financial support provider by the AICTE, New Delhi, India through RPS project scheme and the UGC, New Delhi, India through BSR Ph.D. fellowship to carry out the research work.

REFERENCES

- [1] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.
- [2] Y. Pang and X. Zhang, "Interpolated compressed sensing for 2D multiple slice fast MR imaging," *Ed. Jonathan A. Coles. PLoS ONE*, vol. 8, no. 2, pp. 1–5, 2013.
- [3] Y. Pang, B. Yu, and X. Zhang, "Enhancement of the low resolution image quality using randomly sampled data for multi-slice MR imaging," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 2, pp. 136–144, 2014.
- [4] S. Datta and B. Deka, "Efficient interpolated compressed sensing reconstruction scheme for 3D MRI," *IET Image Processing*, vol. 12, no. 11, p. 2119–2127, 2018.
- [5] ———, "Magnetic resonance image reconstruction using fast interpolated compressed sensing," *Journal of Optics*, vol. 47, no. 2, pp. 154–165, 2017.

- [6] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: sensitivity encoding for fast MRI," *Magnetic Resonance in Medicine*, vol. 42, no. 5, pp. 952 – 962, 1999.
- [7] D. K. Sodickson and W. J. Manning, "Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays," *Magnetic Resonance in Medicine*, vol. 38, no. 4, pp. 591 – 603, 1997.
- [8] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, "Generalized autocalibrating partially parallel acquisitions (GRAPPA)," *Magnetic Resonance in Medicine*, vol. 47, no. 6, pp. 1202 – 1210, 2002.
- [9] P. M. Jakob, M. A. Griswold, R. R. Edelman, and D. K. Sodickson, "AUTO-SMASH: A self-calibrating technique for SMASH imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 7, no. 1, pp. 42–54, 1998.
- [10] M. Lustig and J. Pauly, "SPIRiT: iterative self-consistent parallel imaging reconstruction from arbitrary k-space," *Magnetic Resonance in Medicine*, vol. 64, no. 2, pp. 457–471, 2010.
- [11] D. Liang, B. Liu, J. Wang, and L. Ying, "Accelerating SENSE using compressed sensing," *Magnetic Resonance in Medicine*, vol. 62, no. 6, pp. 1574 – 1584, 2009.
- [12] M. Uecker, P. Lai, M. J. Murphy, P. Virtue, M. Elad, J. M. Pauly, S. S. Vasanawala, and M. Lustig, "ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI: where SENSE meets GRAPPA," *Magnetic Resonance in Medicine*, vol. 71, no. 3, pp. 990 –1001, 2014.
- [13] A. Majumdar and R. K. Ward, "Calibration-less multi-coil MR image reconstruction," *Magnetic Resonance Imaging*, vol. 30, no. 7, pp. 1032 – 1045, 2012.
- [14] P. J. Shin, P. E. Z. Larson, M. A. Ohliger, M. Elad, J. M. Pauly, D. B. Vigneron, and M. Lustig, "Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion," *Magnetic Resonance in Medicine*, vol. 72, pp. 959–970, 2014.
- [15] C. Chen, Y. Li, and J. Huang, "Calibrationless parallel MRI with joint total variation regularization," in *International Conference on Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, ser. LNCS, vol. 8151, 2013, pp. 106–114.
- [16] C. Chen and J. Huang, "Exploiting the wavelet structure in compressed sensing MRI," *Magnetic Resonance Imaging*, vol. 32, pp. 1377–1389, 2014.
- [17] C. Chen, Y. Li, and J. Huang, "Forest sparsity for multi-channel compressive sensing," *IEEE Transactions on Signal Processing*, vol. 62, no. 11, pp. 2803–2813, 2014.
- [18] J. Huang, S. Zhang, and D. N. Metaxas, "Efficient MR image reconstruction for compressed MR imaging," *Medical Image Analysis*, vol. 15, no. 5, pp. 670–679, 2011.
- [19] J. Huang, C. Chen, and L. Axel, "Fast multi-contrast MRI reconstruction," *Magnetic Resonance Imaging*, vol. 32, no. 10, pp. 1344 – 1352, 2014.
- [20] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [21] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

Performance Analysis of BCH and Repetition Codes in Gamma-Gamma Faded FSO Link

Sonali, Nancy Gupta, Abhishek Dixit, and Virander Kumar Jain

Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India

Email: sonali@ee.iitd.ac.in, jop162051@physics.iitd.ac.in, abhishek.dixit@iitd.ac.in, vkjain@ee.iitd.ac.in

Abstract— In this research paper, we have analyzed free space optical (FSO) link performance for intensity modulated/direct detection system. FSO communication link is key to the next generation 5G/10G wireless networks. Generally, FSO link performance is impaired by atmospheric turbulence. Utilizing error correcting codes (ECCs) is one of the mitigation techniques employed to reduce the impact of atmospheric turbulence. We explore Bose-Chaudhuri-Hocquenghem (BCH) and repetition codes for mitigating the effects caused by atmospheric turbulence. We have made the performance comparison in terms of bit error rate (BER) under different turbulence regimes, viz., low, moderate and high regime for both uncoded and coded systems. We evaluate the analytical results and validate them with the simulations. It is concluded that BCH coded system performance is always better than the repetition coded system in all the turbulence regimes. BCH coded system provide a coding gain of 23.4 dB at a high turbulence level which reduces to 17.5 dB and 5 dB at moderate and low turbulence regimes, respectively. However, when we combine the benefits of both the coding schemes, the performance improvement obtained is much higher when compared with both the codes individually. It is, of course, at the cost of an increase in transmission bandwidth requirement.

Keywords— Free space optical (FSO) link, error correcting codes (ECCs), BCH, repetition, bit error rate (BER), coding gain

I. INTRODUCTION

The motivation of the wireless communication system is to cater the growing requirements of high data rate applications such as internet access at low cost, transfer of large files, interactive voice, data and quick deployable system. Existing radio frequency technologies are limited in data rate and bandwidth and pose security issues for mission sensitive data. For these applications, free space optical (FSO) communication is a promising technology [1]. Some of the advantages of the FSO communication system are narrow beam divergence, unlicensed spectrum, security, huge modulation bandwidth and less power requirement [1], [2]. In certain remote areas like hills and mountains where laying of fiber is a difficult and costly affair, an FSO system can be used conveniently. As FSO links are based on line-of-sight technology, an optical carrier with message signal propagates through the atmosphere rather than through optical fiber [2]. These links can be terrestrial links say between two buildings, space links between two satellites or a combination of terrestrial and space links.

In the atmosphere, air masses known as atmospheric turbulence are randomly formed due to the refractive index variation with pressure and temperature. These air masses interact with an optical beam and introduce random

fluctuations in the transmitted optical beam and severely degrade the link performance. It is necessary to mitigate the atmospheric effects to maintain the desired performance level. Various techniques which are used to mitigate the atmospheric effects are error correcting codes (ECCs), appropriate modulation schemes, diversity techniques, etc. [2]. The ECCs can potentially improve the performance in FSO communication. In this paper, we present the effect of ECCs such as Bose-Chaudhuri-Hocquenghem (BCH) code and repetition code on bit error rate (BER) performance of the FSO communication link under different turbulence regimes. ECCs add extra bits (redundancy) to the information bits to make the transmission of data more robust to atmospheric conditions present in the channel [3]. The receiver upon receiving the information sequence removes those redundant bits and extracts the original information sequence transmitted. This technique maintains security and reliability of the network and avoids retransmission of data. ECCs thus provide an effective and reliable FSO communication link. The novelty of our work is in the evaluation of the BCH and the repetition coded FSO link performance individually and in the concatenation, and their comparative study. This kind of work is not yet available in the literature on the wireless optical communication system.

The remainder of the paper is organized as follows. We describe the system and channel models in Section II. Furthermore, we discuss the ECCs (BCH and repetition) and their BER analysis for the FSO link in Section III. Section IV contains the numerical results obtained for coded FSO system. Subsequently, we make a comparative study by presenting simulation and analytical results for both uncoded and coded systems. In Section V, the conclusion of the study is given.

II. SYSTEM AND CHANNEL MODELS

In this section, we first discuss the system model and associated simulation parameters used in this study. Next, we describe Gamma-gamma (GG) distributed channel model.

A. System Model

The system model used in this paper is shown in Fig. 1. The fading channel between the optical source (Laser/LED) and detector (photodiode) pair is considered to be gamma-gamma (G-G) distributed. We have used G-G fading because it can model all the turbulence levels optimally and its statistics match the experimental data when the multiple scattering effects are considered [1]. To mitigate the fading effects we have explored the use of ECCs and employed BCH and repetition code for the same. The encoder is either

BCH/repetition or concatenation of both the schemes depending upon the position of the switches S_1 , S_2 , S_3 and S_4 . Encoded bits are sent through the channel using intensity modulation and direct detection (IM/DD) with on-off keying (OOK) scheme [4]. The received symbols after detection are sent to the respective decoder. After decoding, the decision device decides about transmitted bits. As the received signal is affected by the fading channel and is also corrupted by thermal and shot noises at the receiver, some decisions may go wrong and results in errors. The signal at the input to the decision device is given by

$$\mathbf{Y} = \mathbf{H}\mathfrak{R}\mathbf{X} + \mathbf{I}_{sn} + \mathbf{I}_{th} \quad (1)$$

where \mathbf{Y} represents the received data vector, \mathfrak{R} the responsivity, \mathbf{H} the channel coefficient vector, \mathbf{X} transmitted signal vector, \mathbf{I}_{sn} the shot noise current vector and \mathbf{I}_{th} the thermal noise current vector. The \mathbf{I}_{sn} is both multiplicative and additive white Gaussian noise (WGN) current with [5]

$$\overline{i_{sn}^2} = 2e\mathfrak{R}\mathbf{H}\mathbf{X}\mathbf{B} \quad (2)$$

where e is an electron charge, B the receiver bandwidth and $\overline{i_{sn}^2}$ the mean squared shot noise current variance [5]. In Fig. 1, κ is the constant factor ($\sqrt{2eB}$) in shot noise current. Similarly, i_{th} is an additive white Gaussian noise (AWGN) current with [5]

$$\overline{i_{th}^2} = \frac{4KTB}{R_L} \quad (3)$$

where K is the Boltzmann's constant, T the room temperature, R_L the load resistor and $\overline{i_{th}^2}$ the mean squared thermal noise current variance. Eqs. (1) – (3) are used in determining the BER for the uncoded system later. For the simulations in MATLAB®, the parameters for the system are given in Table I.

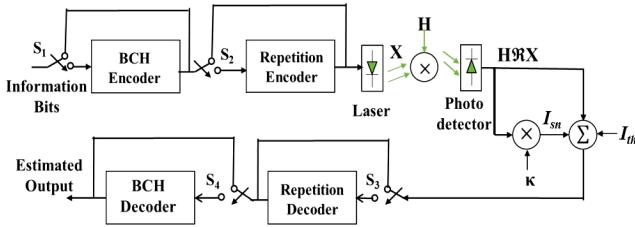


Fig. 1. Block diagram of BCH/repetition coded FSO communication link.

TABLE I. SIMULATION PARAMETERS

Sr. No.	Parameter	Value
1	Data Rate, R_b	1 Gb/s
2	Bandwidth, B	1 GHz
3	Code Rate, R	1/3
4	Link Length, L	2 km
5	Responsivity, \mathfrak{R}	0.8 A/W
6	Receiver Temperature, T	300 K
7	Load Resistance, R_L	50 Ω

B. Channel Model

We can model atmospheric turbulence induced fading by different distributions. For this work, we choose G-G distribution which is an optimum probability distribution function (pdf) to model all the turbulence regimes effectively. It is briefly discussed below.

G-G distribution: Fluctuations in irradiance, which arise due to large and small-sized turbulent eddies, can be modeled by multiplying two independent gamma processes [6]. The resultant multiplication is called G-G distribution which can model weak, moderate and high turbulence regimes optimally [7]. G-G pdf in terms of irradiance h is given as

$$f_H(h) = \frac{2(\alpha\beta)^{(\alpha+\beta)/2}}{\Gamma(\alpha)\Gamma(\beta)} h^{(\alpha+\beta)/2-1} K_{\alpha-\beta}(2\sqrt{\alpha\beta h}) \quad (4)$$

where α and β denote the shape and scale parameter of G-G pdf, which is the measure of small-scale and large-scale eddies experienced by the optical beam, Γ is the gamma function and $K_{\alpha-\beta}$ the modified Bessel function of the second kind and order, $(\alpha-\beta)$. We can model the strength of turbulence by the inclusion of parameters α and β in the characterization of irradiance fluctuations. The irradiance fluctuations are measured in terms of scintillation index which is given by

$$\sigma_{SI}^2 = \exp \left[\frac{0.49\sigma_l^2}{(1+1.11\sigma_l^{12/5})^{7/6}} + \frac{0.51\sigma_l^2}{(1+0.69\sigma_l^{12/5})^{5/6}} \right] - 1 \quad (5)$$

where σ_l^2 represents the Rytov variance [1], [2] and is given by

$$\sigma_l^2 = 1.23C_n^2 k^{7/6} L^{11/6} \quad (6)$$

Here, $k = 2\pi/\lambda$, λ is the wavelength, L the propagation length and C_n^2 the refractive index structure parameter [4]. The value of the parameter C_n^2 depends upon the time of the day and the height at which turbulence is considered. Further, σ_l^2 decides the value of parameters α and β as

$$\alpha = \left[\exp \left(\frac{0.49\sigma_l^2}{(1+1.11\sigma_l^{12/5})^{7/6}} \right) - 1 \right]^{-1} \quad (7)$$

and

$$\beta = \left[\exp \left(\frac{0.51\sigma_l^2}{(1+0.69\sigma_l^{12/5})^{5/6}} \right) - 1 \right]^{-1} \quad (8)$$

The G-G distribution parameters used to model high to low turbulence conditions are given in Table II.

TABLE II. G-G DISTRIBUTION PARAMETERS

Turbulence Regime	α	β	σ_t^2
High	4.2	1.4	3.5
Moderate	4.0	1.9	1.6
Low	11.6	10.1	0.2

III. ERROR CORRECTING CODES

In this section, we give a brief description of the encoding and decoding strategies of the ECCs used in this work.

A. BCH Code

It is a type of binary cyclic ECCs which is defined over the Galois field and can correct multiple bits in error [8]. It is a linear block code in which redundancy is added to the input signal. We consider BCH code in which the k information bits are converted into n bit codeword with $n-k$ bits as redundant bits. Code rate (R) which is given by k/n represents the redundancy. In our work, we have taken R as 1/3. The basic block diagram for the BCH encoding/decoding is shown in Fig. 2.

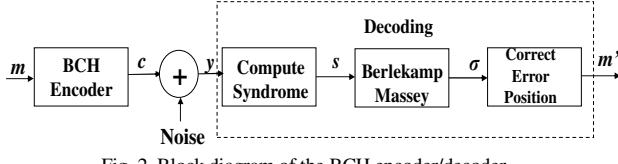


Fig. 2. Block diagram of the BCH encoder/decoder.

The message bits m are encoded using BCH encoder and sent through the FSO channel. The received message bits y are sent to the BCH decoder and are estimated as m' . The decoding of the BCH code is done in three steps as shown in Fig. 2. First, the syndrome is computed, then the Berlekamp-Massey algorithm is used to generate the error location polynomial from syndrome polynomial. Next step is to get error locations by finding the roots of error location polynomial. Errors are corrected and estimated information word is generated after decoding [8]. This category of codes permits high flexibility which allows control over the block length (n) and error patterns (t) which can be easily corrected. This leads to a mathematical relationship between n and t which is given as

$$\frac{t}{1-R} \geq \frac{n}{\log_2(n+1)} \quad (9)$$

Thus for a required t , we can choose optimum n using the above equation. The probability of error for BCH code can be approximated using the generalized BER formulation for linear block codes and is given by [9]

$$P_{e,BCH} \cong \frac{1}{n} \sum_{i=t+1}^n i \binom{n}{i} p_e^i (1-p_e)^{n-i} \quad (10)$$

where p_e is the average BER for coded transmission before decoding, and is obtained as

$$P_e = \int_0^\infty \frac{1}{2} erfc \left(\frac{\Re P_t h}{\sqrt{2}(\sigma_0 + \sigma_1)} \right) \times f_H(h) dh \quad (11)$$

In the above equation, P_t is the transmitted power, $f_H(h)$ the channel pdf, $\sigma_o^2 = \bar{i}_{th}^2$, $\sigma_i^2 = \bar{i}_{sn}^2 + \bar{i}_{th}^2$, erfc(.) the complementary error function and h denotes the fading coefficient. With equiprobable bit ‘1’ and ‘0’, the average transmitted power will be $P_t/2$.

B. Repetition Code

The simplest coding technique known is the repetition code. Its encoding procedure involves, simply repeating the information bits N times resulting in $(N, 1)$ block code and the decoding is done by majority logic decoding [10]. It has only two codewords - all ‘0’ codeword and all‘1’ codeword. We declare bit ‘0’ is received if ‘0’ is received more than $N/2$ times else decision is made in favor of another bit. For repetition coding, the code rate is given by $1/N$. When N is odd, this type of code can correct all error patterns of weight given by [11]

$$t \leq \frac{N-1}{2} \quad (12)$$

The probability of error for repetition code can be approximated using the generalized BER formulation for maximum likelihood decoding and is given by [12]

$$P_{e,Repetition} = \sum_{j=N/2}^N \binom{N}{j} p_e^j (1-p_e)^{n-j} \quad (13)$$

where p_e is directly taken from Eq. (11) and substituted in Eq. (13) to obtain the probability of error for repetition coding.

IV. NUMERICAL RESULTS

In this section, we evaluate the BER performance of an FSO link with ECCs (BCH/repetition and both) under different turbulence regimes. For both the codes, simulations are done at the code rate of 1/3. In this subsection, we compare the performance of an uncoded and coded FSO systems in all the three turbulence regimes and evaluate the coding gains. Variation of BER with average transmitted power for BCH ($n = 255$, $k = 87$) and repetition (1, 3) codes are given in Figs. 3, 4 and 5 under high, moderate and low turbulence regimes, respectively. For all the simulation results, their corresponding analytical results are plotted from Eqs. (10) – (13). The analytical results almost match with the simulation results in all the turbulence regimes. From Subsection III-A, for BCH code, we obtain $t \geq 21$, but for $n = 255$ and $k = 87$, we have $t = 26$ [13]. Corresponding to $t = 26$, the analytical BER is computed using Eq. (10) and is then compared with the results obtained by simulations. The coding gain for this system is evaluated as

$$\text{Coding Gain(dB)} = 10 \log_{10} \left(\frac{P_t(\text{uncoded system})}{P_t(\text{coded system})} \right) \quad (14)$$

The coding gain is the reduction in the power required by the coded system to achieve the same BER performance as

an uncoded system. In this work, coding gain is evaluated at a particular BER value of 10^{-5} as is used in [14]. It is shown in Fig. 3 that at this BER, we get coding gain of 23.4 dB and 11.8 dB for BCH code and repetition code, respectively over the uncoded system. Coding gain is 27.8 dB when we concatenate both the codes. Thus, we conclude that the BCH coded system provides much better performance than the repetition code and to achieve more improvement, we can combine the benefits of both the coding schemes by concatenating them. As far as ease of implementation is concerned, repetition code instead of BCH code can still be preferred despite lower coding gain.

It is shown in Fig. 4 that coding gain reduces when we move from a higher level of turbulence to a moderate level. For a moderate turbulence regime, at a BER of 10^{-5} , we obtain the coding gain of 21.1 dB, 17.5 dB and 8.5 dB for BCH code concatenated with repetition code, BCH and repetition code, respectively. It implies an advantage of 9 dB for BCH code over the repetition code and 3.6 dB for concatenated code over BCH code.

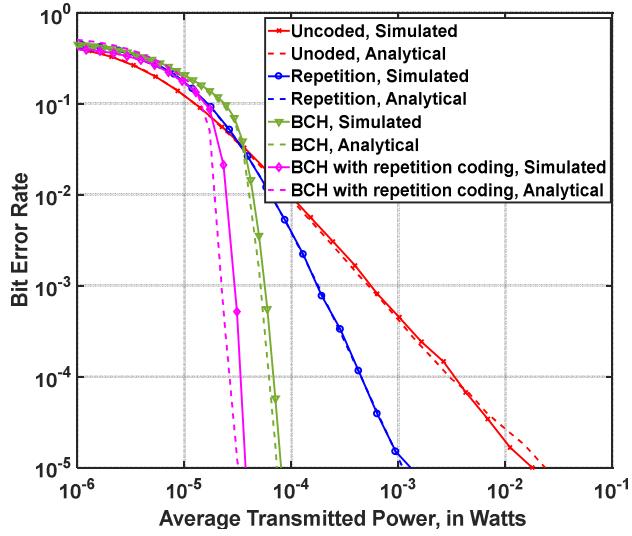


Fig. 3. Analytical and simulation results for high turbulence ($\sigma_i^2 = 3.5$).

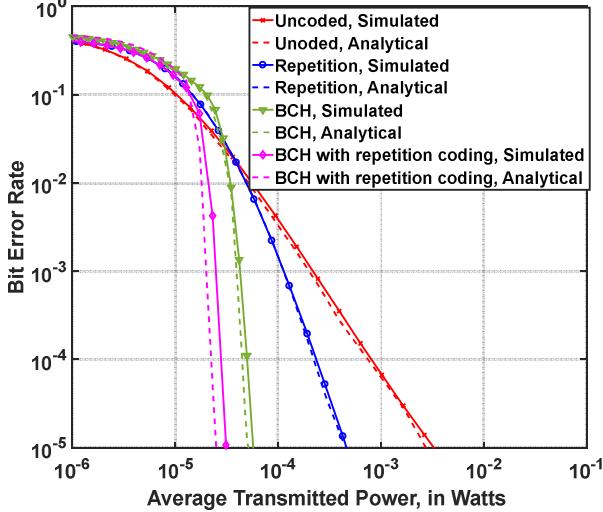


Fig. 4. Analytical and simulation results for moderate turbulence ($\sigma_i^2 = 1.6$).

In Fig. 5, the performance of a coded FSO system against an uncoded system under a low turbulence regime is shown. It is observed from this figure that the target BER is achieved at the much lower value of power than in other turbulence regimes (target BER is achieved at a power level of less than $100 \mu\text{W}$). At a BER of 10^{-5} , we obtain the coding gain of 8.3 dB, 5 dB and 0.3 dB for BCH code concatenated with repetition code, BCH code and repetition code, respectively. The advantage of BCH code over the repetition code further reduces to 4.7 dB and it is 3.3 dB for concatenated code over BCH code.

It is seen from Figs. (3) – (5) that as the level of turbulence reduces, coding gain also reduces for both the BCH and repetition codes. Further, the advantage of the concatenated BCH repetition code over the repetition code is 16 dB, 12.6 dB, and 8 dB in high, moderate and low turbulence regime, respectively. Coding gain reduces with the decreasing level of turbulence because there is less scope of improvement in low turbulence regime because of less errors whereas high turbulence level incurs more errors and scope of improvement is much higher resulting in higher coding gain. These results are given in the tabular form in Table III.

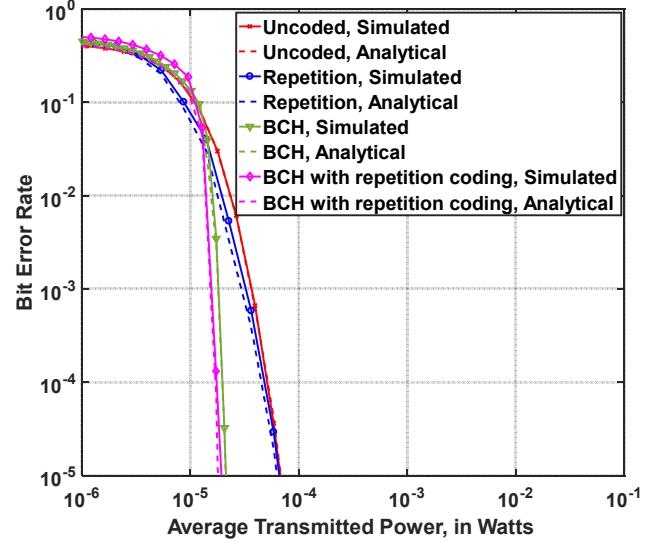


Fig. 5. Analytical and simulation results for low turbulence ($\sigma_i^2 = 0.2$).

TABLE III. CODING GAIN FOR DIFFERENT TURBULENCE REGIMES

Error Correcting Coding Scheme	Coding Gain (dB)		
	High Turbulence	Moderate Turbulence	Low Turbulence
BCH and repetition	27.8	21.1	8.3
BCH	23.4	17.5	5
Repetition	11.8	8.5	0.3

It is therefore concluded that coding improves the error performance of the FSO link in all the turbulence regimes.

Both BCH and repetition codes have resulted in highest performance gain in much-affected turbulence regime, i.e., high turbulence regime, then in moderate and lowest in low turbulence regime. Further, as the level of turbulence increases, the performance advantage of BCH code over repetition code also increases.

V. CONCLUSION

A significant improvement in the performance implying high coding gain is obtained when ECCs are applied in the FSO link. It is seen that BCH code gives much higher coding gain as compared to repetition code but at the cost of complex encoding and decoding strategy. On the other hand, the concatenation of both the codes provides much higher coding gain with almost similar complexity involved as with BCH code alone. This is true for all the turbulence regimes, viz., low, moderate, and high. Maximum coding gain of 27.8 dB is achieved in a high turbulence regime with concatenated BCH and repetition code. Coding gain reduces to 21.1 dB and 8.3 dB in moderate and low turbulence regimes, respectively. This work quantitatively highlights the significance of ECCs in the performance improvement of a typical FSO link.

ACKNOWLEDGMENT

This work was supported by the Department of Electronics and Information Technology (DeitY), Govt. of India under the project, Visvesvaraya Ph.D. Scheme for Electronics and IT at IIT Delhi.

REFERENCES

- [1] Z. Ghassemlooy and W. O. Popoola, *Terrestrial Free-Space Optical Communications*, CRC Press, Florida, 2010.
- [2] H. Kaushal, V. K. Jain and S. Kar, *Free Space Optical Communication*, Springer, India, 2017.
- [3] I. Djordjevic, W. Ryan and B. Vasic, *Coding for Optical Channels*, Springer, New York, 2010.
- [4] Z. Ghassemlooy, W.O. Popoola, S. Rajbhandari, M. Amiri and S. Hashemi, "A synopsis of modulation techniques for wireless infrared communication," in *9th IEEE, International Conference on Transparent Optical Networks*, Rome, pp. 1–6, July 2007.
- [5] G. P. Agrawal, *Fiber-Optic Communication Systems*, vol. 2, John Wiley & Sons, New York, 2012.
- [6] S. Garg, A. Dixit and V. K. Jain, "Performance analysis of STBC-FSO communication system in different turbulence regimes," in *Advances in Communication, Devices and Networking*, Springer, Singapore, pp. 409-417, Feb. 2018.
- [7] K. Prabu, D. S. Kumar and T. Srinivas, "Performance analysis of FSO links under strong atmospheric turbulence conditions using various modulation schemes," *Optik-International Journal for Light and Electron Optics*, vol. 125, no. 19, pp. 5573-5581, Oct. 2014.
- [8] C. Ding, "Parameters of several classes of BCH codes," *IEEE Transactions on Information*, vol. 61, no. 10, pp. 5322-5330, Oct. 2015.
- [9] A. Al-Barak, A. Al-Sherbaz, T. Kanakis and R. Crockett, "Enhancing BER performance limit of BCH and RS codes using multipath diversity," *Computers*, vol. 6, no. 2, Jun. 2017.
- [10] S. K. Hanna and S. E. Rouayheb, "Guess & check codes for deletions, insertions, and synchronization," in *IEEE Transactions on Information Theory*. doi: 10.1109/TIT.2018.2841936.
- [11] E. Modiano. (2018, June). Communication Systems Engineering, M.I.T. Open Course Ware [Online]. Available: https://ocw.mit.edu/courses/aeronautics-and-astronautics/16-36-communication-systems-engineering-spring-2009/lecturenotes/MIT_16_36s09_lec13_14.pdf
- [12] D. J. Mac Kay, *Information Theory, Inference and Learning Algorithms*, Cambridge university press, 2003.
- [13] E. Fujiwara, *Code Design for Dependable Systems: Theory and Practical Applications*, John Wiley & Sons, New Jersey, 2006.
- [14] G. D. Forney. (2018, June). Principles of Digital Communication II, M.I.T. Open Course Ware [Online]. Available: <http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-451Spring-2005/Course Home/index.htm>

Towards the Exact Rate Memory Tradeoff in Coded Caching

Vijith Kumar K P
 Department of EEE
 IIT Guwahati, INDIA
 Email: vijith@iitg.ernet.in

Brijesh Kumar Rai
 Department of EEE
 IIT Guwahati, INDIA
 Email: bkrai@iitg.ernet.in

Tony Jacob
 Department of EEE
 IIT Guwahati, INDIA
 Email: tonyj@iitg.ernet.in

Abstract—Caching plays an important role in improving internet performance by keeping a fraction of the files closer to the end user. The peak data traffic in the network can be significantly reduced by proper utilization of caching. Recent studies have shown that coded caching does help in further reducing the data traffic over uncoded caching. In this paper, we consider the problem of the exact rate memory tradeoff in coded caching. For the $(3,3)$ canonical cache network, a new caching scheme to achieve the memory rate pair $(5/3, 1/2)$ is introduced. This scheme is further extended to the $(4,4)$ canonical cache network, to achieve the memory rate pair $(11/4, 1/3)$. We prove the optimality of both the proposed schemes by deriving new lower bounds and thus partially characterizing the exact rate memory tradeoff in coded caching.

Index Terms—Coded caching, coded pre-fetching, exact rate memory tradeoff.

I. INTRODUCTION

Over the last decades, there has been an exponential growth in data traffic over the internet. One way to keep up with this is to use caching, where, rather than serving the data from the origin, it is served from a local source near to the user. Caching aims to reduce the peak time data traffic by keeping a fraction of the files close to the end users. The fundamental issue in caching is to decide what to place in the cache and accordingly what to deliver to fulfil the users' demand in such a way that the network experiences minimum peak data traffic. It is worth noting that the problem of caching has been extensively studied in the area of computer organization, where, typically, there is a single cache. There have been various interesting results in this area, but they do not adequately apply in the case of a network of caches.

In this paper, we consider the (N, K) canonical cache network presented in [1], as shown in Fig. 1. In this network, there is a server which has N files, $\{W_1, \dots, W_N\}$, each of size F bits. There are K users who are connected to the server through an error free shared link. Each user has an isolated cache memory of size MF bits, where $M \in [0, N]$. Let U_k denote the user k and Z_k denote its cache. The caching procedure consists of two phases, the placement phase and the delivery phase. The placement takes place in the off-peak time when, the server places MF bits into each user's cache without any knowledge about user's future requests. The delivery phase takes place during the peak traffic time when, each

user communicates their requests with the server. The user's requests can be represented as a vector $\mathbf{d} = \{W_{d_1}, \dots, W_{d_K}\}$, where W_{d_k} represent the file requested by the user k in the demand \mathbf{d} . Once the server is informed of these requests, the server transmits a signal $X_{\mathbf{d}}$ of size RF bits, for some real number R . The quantity RF is called the load experienced by the shared link and R is called the rate. A *memory rate pair* (M, R) is said to be achievable, if each user can recover the requested file from the received packet $X_{\mathbf{d}}$ with the help of the cached contents.

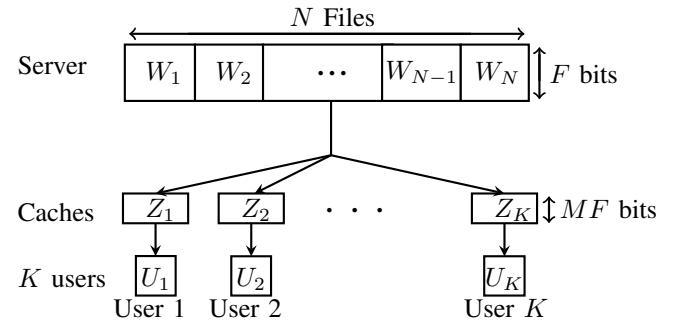


Fig. 1: The canonical cache network considered in [1].

In their seminal work [1], Maddah-Ali and Niesen analysed the (N, K) canonical cache network and showed that coded caching does help in reducing the peak data rate over uncoded caching. In [1], the authors introduced a coded caching scheme and through cut set arguments proved that the scheme is within a multiplicative gap of 12 from the optimal scheme. Several improvements to this result have been presented in [3], [8]–[10], with the current best multiplicative gap of 2 in [10]. This notion of coded caching is then extended to the case of decentralized cache networks [11], hierarchical cache networks [12], cache networks with non-uniform demands [13], cache networks with multi-servers [14], online caching placement [15] etc.

The scheme, presented in [1], has an uncoded placement phase and a coded delivery phase. This scheme was proved to be optimal in the case of uncoded placement for the (N, K) canonical cache network where number of files is at least the number of users, i.e., $N \geq K$ [6]. Yu et al., in a surprising result [7], proved that a modified version of this scheme is

Caching Scheme	Placement Phase	Cache Size, M	Rate, R	Matching Lower Bound
Chen et al. [2]	Coded pre-fetching	$\frac{1}{K}$	$N\left(1 - \frac{1}{K}\right)$	Cut set bound
Amiri and Gunduz [3]	Coded pre-fetching	$\frac{N-1}{K}$	$N\left(1 - \frac{N}{2K}\right)$	No matching lower bound
Gómez-Vilardebó [4]	Coded pre-fetching	$\frac{N}{Kq}$, where $q \in \{1, 2, \dots, N\}$	$N\left(1 - \frac{N+1}{K(q+1)}\right)$	Matching lower bound for $K = N$ and $M = 1/(N-1)$
Tian and Chen [5]	Coded pre-fetching	$\frac{t[(N-1)t + K - N]}{K(K-1)}$, where $t \in \{0, 1, \dots, N\}$	$N\left(1 - \frac{t}{K}\right)$	No matching lower bound
Maddah-Ali and Niesen [1]	Uncoded pre-fetching	$\frac{Nr}{K}$, where $r \in \{0, 1, \dots, K\}$	$\binom{K-r}{1+r}$	Matching lower bound for uncoded pre-fetching when $N \geq K$ [6]
Yu et al. [7]	Uncoded pre-fetching	$\frac{Nr}{K}$, where $r \in \{0, 1, \dots, K\}$	$\frac{\binom{K}{r+1} - \binom{K-N}{r+1}}{\binom{K}{r}}$	Matching lower bound for uncoded pre-fetching

TABLE I: Previous works in coded caching.

exact optimal for the general (N, K) canonical cache network when placement phase is restricted to be uncoded. The authors of [1] also considered the problem of optimal schemes which permit coding in the placement phase as well as in the delivery phase. For the $(2, 2)$ canonical cache network, they introduced a scheme and derived a matching lower bound to provide a complete characterization of the exact rate memory tradeoff, depicted in Fig. 2. In [16], Tian proved that the scheme presented in [7] completely characterizes the exact rate memory tradeoff for the $(N, 2)$ canonical cache network where number of files is at least three, i.e., $N \geq 3$.

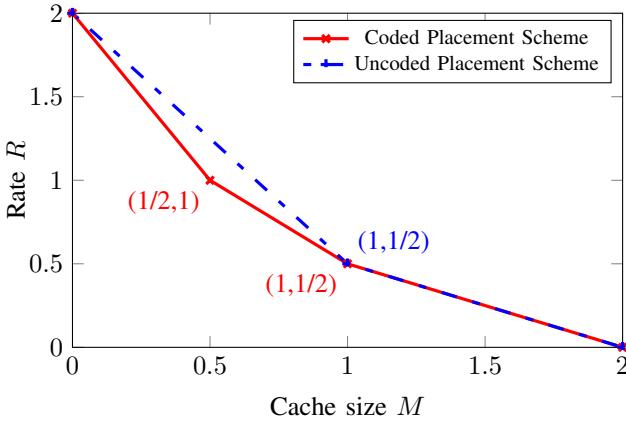


Fig. 2: Exact (M, R) tradeoff for the $(2,2)$ cache network.

A natural question in this context is that of characterizing the exact rate memory tradeoff for the general (N, K) canonical cache network where each file is requested by at least one user. Several schemes were presented in [2]–[5], [7] to improve the achievable rate obtained in [1] and these results are summarised in TABLE I. It can be noted that the problem of characterizing the exact rate memory tradeoff in the case of uncoded placement is solved in [7]. In this paper we propose new schemes which leads to an improvement on characterization of the exact rate memory tardeoff for the $(3, 3)$

and $(4, 4)$ canonical cache networks.

The rest of the paper is organized as follows. In Section II, we introduce a scheme for the $(3, 3)$ canonical cache network for the memory rate pair $(5/3, 1/2)$ and derive a matching lower bound. We further extend this to the $(4, 4)$ canonical cache network and demonstrate its optimality in Section III. We conclude the paper with few remarks in Section IV.

II. THE $(3, 3)$ CANONICAL CACHE NETWORK

In this section, we consider the $(3, 3)$ canonical cache network. In this network, the server has three files $\{A, B, C\}$, each of size F bits and the three users are connected to the server through a common shared link. For this network Tian, in [17], derived a lower bound, as depicted in Fig. 3. From the known achievable rate curve (in Fig. 3), it is clear that, till now there exists no scheme to achieve the memory rate pairs $(2/3, 4/3)$, $(7/6, 5/6)$ and $(5/3, 1/2)$. In this section, we present a new caching scheme to achieve the optimum memory rate pair $(5/3, 1/2)$.

During the placement phase, we split each file into 6 disjoint subfiles of size $F/6$ bits. Let the subfiles corresponding to the files A , B , and C be represented as,

$$\begin{aligned} A &= \{A_1, A_2, A_3, A_4, A_5, A_6\}, \\ B &= \{B_1, B_2, B_3, B_4, B_5, B_6\}, \\ C &= \{C_1, C_2, C_3, C_4, C_5, C_6\}. \end{aligned}$$

A set of coded packets is placed into each user's cache. The users' cache contents are as shown in TABLE II.

Due to the symmetry in the cached contents, instead of considering each demand separately, we consider a general demand to analyse the delivery phase. Consider a demand $\mathbf{d} = \{P, Q, R\}$, where P represents the file requested by the user 1, Q represents the file requested by the user 2 and R represents the file requested by the user 3, which all assumed to be distinct. In response to this demand the server broadcasts a set of packets,

$$X_{\mathbf{d}} = \{P_4 + Q_2, P_6 + R_1, Q_5 + R_3\}.$$

Proof for lemma 2.

$$\begin{aligned}
3M + 6R &\geq 3H(Z_1) + 6H(X_{\{A,B,C\}}) \\
&\geq H(Z_1, X_{\{A,B,C\}}, X_{\{B,C,A\}}) + H(Z_2, X_{\{A,B,C\}}, X_{\{B,A,C\}}) + H(Z_3, X_{\{C,A,B\}}, X_{\{B,C,A\}}) \\
&\stackrel{(a)}{=} H(A, B, Z_1, X_{\{A,B,C\}}, X_{\{B,C,A\}}) + H(A, B, Z_2, X_{\{A,B,C\}}, X_{\{B,A,C\}}) + H(A, B, Z_3, X_{\{C,A,B\}}, X_{\{B,C,A\}}) \\
&\stackrel{(b)}{\geq} H(A, B, X_{\{A,B,C\}}) + H(A, B, Z_1, Z_2, X_{\{A,B,C\}}, X_{\{B,A,C\}}, X_{\{B,C,A\}}) + H(A, B, Z_3, X_{\{C,A,B\}}, X_{\{B,C,A\}}) \\
&\stackrel{(c)}{=} H(A, B, X_{\{A,B,C\}}) + H(A, B, C) + H(A, B, Z_3, X_{\{C,A,B\}}, X_{\{B,C,A\}}) \\
&\stackrel{(b)}{\geq} H(A, B, C) + H(A, B) + H(A, B, Z_3, X_{\{C,A,B\}}, X_{\{B,C,A\}}, X_{\{A,B,C\}}) \\
&\stackrel{(a)}{=} H(A, B, C) + H(A, B) + H(A, B, C, Z_3, X_{\{C,A,B\}}, X_{\{B,C,A\}}, X_{\{A,B,C\}}) \\
&\stackrel{(c)}{=} H(A, B, C) + H(A, B) + H(A, B, C) \\
&= 3 + 2 + 3 = 8,
\end{aligned}$$

where (a) follows from (1), (b) follows from submodularity of entropy and (c) follows from (2).

User 1, Z_1	User 2, Z_2	User 3, Z_3
A_1	A_3	A_5
A_2	A_4	A_6
B_1	B_3	B_5
B_2	B_4	B_6
C_1	C_3	C_5
C_2	C_4	C_6
$A_3 + A_5$	$A_1 + A_6$	$A_2 + A_4$
$B_3 + B_5$	$B_1 + B_6$	$B_2 + B_4$
$C_3 + C_5$	$C_1 + C_6$	$C_2 + C_4$
$A_3 + B_3 + C_3$	$A_1 + B_1 + C_1$	$A_2 + B_2 + C_2$

TABLE II: Cache content for the (3, 3) cache network.

In this demand, the user 1 requested for the subfiles $\{P_1, \dots, P_6\}$. The subfiles P_1 and P_2 are already available at the user 1's cache memory. The user 1 also has subfiles $\{Q_1, Q_2, R_1, R_2\}$ cached in its cache memory. By XORing these subfiles with the received packets $P_4 + Q_2$ and $P_6 + R_1$, the user 1 obtains the subfiles P_4 and P_6 . The user 1, upon receiving the packet $Q_5 + R_3$, evaluates the function $Q_3 + R_3$ by XORing the cached function $Q_3 + Q_5$ with it. Now, the user 1 obtains the subfile P_3 by XORing the evaluated function $Q_3 + R_3$ with the cached function $P_3 + Q_3 + R_3$. Once the subfile P_3 is recovered, the user 1 obtains the subfile P_5 by XORing the subfile P_3 with cached function $P_3 + P_5$. Similarly, the user 2 and the user 3 can obtain the requested files Q and R respectively.

It can be noted that in response to the demand, the server broadcasts three packets, each of size $\frac{F}{6}$ bits. Therefore, the corresponding rate experienced by the common link is $R = 1/2$. Thus:

Lemma 1. For the (3, 3) canonical cache network, there exists a caching scheme to achieve the memory rate pair $(5/3, 1/2)$.

In [17], Tian presented a lower bound for the (3, 3) canonical cache network:

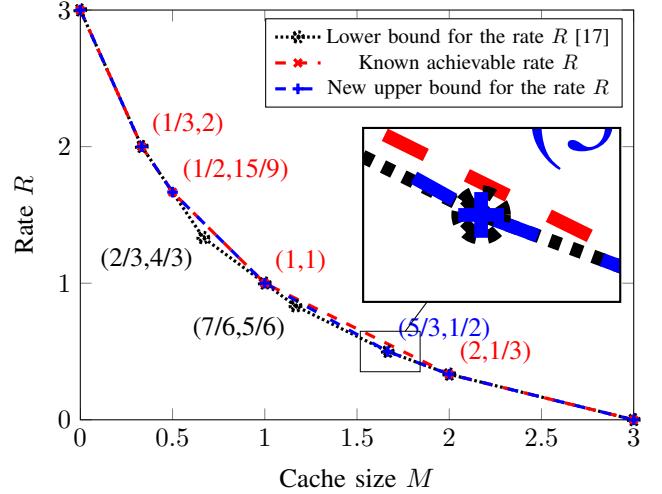


Fig. 3: Rate memory tradeoff for the (3, 3) cache network.

Lemma 2. (Tian [17]) For the (3, 3) canonical cache network achievable memory rate pairs (M, R) must satisfy

$$3M + 6R \geq 8.$$

In [17], Tian use the computational approach to prove this constraint. Here we give a mathematical proof for the same. The following identities are used to prove lemma 2 and lemma 4.

$$H(W_{d_k}, Z_k, X_d) = H(Z_k, X_d) \quad (1)$$

$$H(W_{[N]}, Z_k, X_d) = H(W_{[N]}) \quad (2)$$

Proof of these identities are given in the appendix. For the (3, 3) canonical cache network, we can achieve the memory rate pair $(5/3, 1/2)$ by using the scheme presented above and the memory rate pair $(2, 1/3)$ by using the scheme presented

User 1, Z_1	User 2, Z_2	User 3, Z_3	User 4, Z_4
A_1	A_1	A_3	A_5
A_2	A_2	A_4	A_6
A_3	A_7	A_7	A_8
A_4	A_8	A_9	A_9
A_5	A_{10}	A_{10}	A_{11}
A_6	A_{11}	A_{12}	A_{12}
B_1	B_1	B_3	B_5
B_2	B_2	B_4	B_6
B_3	B_7	B_7	B_8
B_4	B_8	B_9	B_9
B_5	B_{10}	B_{10}	B_{11}
B_6	B_{11}	B_{12}	B_{12}
C_1	C_1	C_3	C_1
C_2	C_2	C_4	C_6
C_3	C_7	C_7	C_8
C_4	C_8	C_9	C_9
C_5	C_{10}	C_{10}	C_{11}
C_6	C_{11}	C_{12}	C_{12}
D_1	D_1	D_3	D_5
D_2	D_2	D_4	D_6
D_3	D_7	D_7	D_8
D_4	D_8	D_9	D_9
D_5	D_{10}	D_{10}	D_{11}
D_6	D_{11}	D_{12}	D_{12}
$A_7 + A_8$	$A_3 + A_5$	$A_1 + A_6$	$A_2 + A_4$
$A_8 + A_9$	$A_5 + A_{12}$	$A_6 + A_{11}$	$A_4 + A_{10}$
$B_7 + B_8$	$B_3 + B_5$	$B_1 + B_6$	$B_2 + B_4$
$B_8 + B_9$	$B_5 + B_{12}$	$B_6 + A_{11}$	$B_4 + B_{10}$
$C_7 + C_8$	$C_3 + C_5$	$C_1 + C_6$	$C_2 + C_4$
$C_8 + C_9$	$C_5 + C_{12}$	$C_6 + A_{11}$	$C_4 + C_{10}$
$D_7 + D_8$	$D_3 + D_5$	$D_1 + D_6$	$D_2 + D_4$
$D_8 + D_9$	$D_5 + D_{12}$	$D_6 + D_{11}$	$D_4 + D_{10}$
$A_7 + B_7 + C_7 + D_7$	$A_3 + B_3 + C_3 + D_3$	$A_1 + B_1 + C_1 + D_1$	$A_2 + B_2 + C_2 + D_2$

TABLE III: Cache content for the $(4, 4)$ cache network.

in [1]. By memory sharing these schemes, we can achieve memory rate pairs satisfying,

$$R(M) = \frac{4}{3} - \frac{1}{2}M, \quad (3)$$

where $M \in [5/3, 2]$. From lemma 2 we have,

$$\begin{aligned} R &\geq \frac{8}{6} - \frac{3}{6}M \\ &= \frac{4}{3} - \frac{1}{2}M = R(M). \end{aligned}$$

We know that the scheme presented in [1] achieves the cut set bound for the $(3, 3)$ canonical cache network when the cache size $M \in [2, 3]$. This completes the characterization of exact rate memory tradeoff for cache size $M \in [5/3, 3]$. The result is summarized in Theorem 1.

Theorem 1. *For the $(3, 3)$ canonical cache network, where each user has a cache of size $M \in [5/3, 3]$, the exact rate memory tradeoff is achieved by memory sharing the proposed code with Maddah-Ali and Niesen's code presented in [1].*

III. THE $(4,4)$ CANONICAL CACHE NETWORK

In this section, we extend our caching scheme for the $(4, 4)$ canonical cache network. In this network, the server has

four files, $\{A, B, C, D\}$, each of size F bits. Four users are connected to the server through a common shared link. Each user has a cache of size $M = 11/4$. During the placement phase we first split each file into 12 disjoint subfiles, each of size $F/12$. Let the subfiles of files A, B, C and D be,

$$\begin{aligned} A &= \{A_1, A_2, \dots, A_{12}\}, \\ B &= \{B_1, B_2, \dots, B_{12}\}, \\ C &= \{C_1, C_2, \dots, C_{12}\}, \\ D &= \{D_1, D_2, \dots, D_{12}\}. \end{aligned}$$

A set of coded packets is placed into each user's cache. The users' cache contents are as shown in TABLE III.

Due to the symmetry in the cached contents, instead of considering each demand separately, we consider a general demand to analyse the delivery phase. Consider a demand $\mathbf{d} = \{P, Q, R, S\}$, where P represents the file requested by the user 1, Q represents the file requested by the user 2, R represents the file requested by the user 3 and S represents the file requested by the user 4. In response to this demand,

the server broadcasts a set of packets,

$$X_{\{P,Q,R,S\}} = \begin{cases} P_{10} + Q_4 + R_2, & P_{11} + Q_6 + S_1, \\ P_{12} + R_5 + S_3, & Q_9 + R_8 + S_7 \end{cases}.$$

In this demand, the user 1 requested for the subfiles $\{P_1, \dots, P_{12}\}$. The subfiles $\{P_1, \dots, P_6\}$ are already available at the user 1's cache memory. The user 1 also has subfiles $\{Q_1, Q_2, Q_3, Q_4, Q_5, Q_6\}$, $\{R_1, R_2, R_3, R_4, R_5, R_6\}$ and $\{S_1, S_2, S_3, S_4, S_5, S_6\}$. By XORing these subfiles with the received packets $P_{10} + Q_4 + R_2$, $P_{11} + Q_6 + S_1$ and $P_{12} + R_5 + S_3$, the user 1 obtain the subfiles P_{10} , P_{11} and P_{12} , respectively. The user 1, upon receiving the packet $Q_9 + R_8 + S_7$, evaluates:

$$Q_7 + Q_9, Q_7 + R_7 + S_7, \text{ and } P_7 + P_9,$$

where $Q_7 + Q_9$ is computed by XORing the cached contents $Q_7 + Q_8$ and $Q_8 + Q_9$, $Q_7 + R_7 + S_7$ is computed by XORing evaluated function $Q_7 + Q_9$ and cached content $R_7 + R_8$ with the received packet $Q_9 + R_8 + S_7$, $P_7 + P_9$ is computed by XORing the cached contents $P_7 + P_8$ and $P_8 + P_9$. The subfile P_7 is obtained by XORing the cached content $P_7 + Q_7 + R_7 + S_7$ with evaluated function $Q_7 + R_7 + S_7$. Once the subfile P_7 is obtained, the user 1 obtains the subfiles P_8 and P_9 by XORing the subfile P_7 with the cached content $P_7 + P_8$ and evaluated function $P_7 + P_9$, respectively. Similarly the user 2, the user 3 and the user 4 can obtain the requested files Q , R and S respectively.

It can be noted that the server broadcasts four packets, each of size $F/12$ bits, in response to this demand. Therefore, the corresponding rate experienced by the common link is $R = 1/3$. Thus:

Lemma 3. *For the (4, 4) canonical cache network, there exist a caching scheme to achieve the memory rate pair $(11/4, 1/3)$.*

For the (4, 4) canonical cache network we have the following lemma.

Lemma 4. *For the (4, 4) canonical cache network achievable memory rate pairs (M, R) must satisfy,*

$$4M + 12R \geq 15. \quad (4)$$

The proof of lemma 4 is given at the top of next page. For the (4, 4) canonical cache network we can achieve the memory rate pair $(11/4, 1/3)$ by using the scheme presented above and the memory rate pair $(3, 1/4)$ by using the scheme presented in [1]. By memory sharing these schemes we can achieve memory rate pairs satisfying,

$$R(M) = \frac{5}{4} - \frac{1}{3}M,$$

where $M \in [11/4, 3]$. From lemma 4, we have,

$$\begin{aligned} R &\geq \frac{15 - 4M}{12} \\ &= \frac{5}{4} - \frac{1}{3}M = R(M). \end{aligned}$$

We know that the scheme presented in [1] achieves the cut set bound for the (4, 4) canonical cache network when the cache size $M \in [3, 4]$. This completes the characterization of exact rate memory tradeoff in the region $M \in [11/4, 4]$. The result is summarized in Theorem 2.

Theorem 2. *For the (4, 4) canonical cache network, where each user has a cache of size $M \in [11/4, 4]$, the exact rate memory tardeoff is achieved by memory sharing the proposed code with Maddah-Ali and Niesen's code presented in [1].*

IV. CONCLUSIONS AND REMARKS

In this paper, we presented a new caching scheme to achieve the optimal memory rate pair $(5/3, 1/2)$ for the (3, 3) canonical cache network for the demand where all the users request for distinct files. We extended this scheme for the (4, 4) canonical cache network to achieve the memory rate pair $(11/4, 1/3)$ for the demand where all the users request for distinct files. Further, we proved that this memory rate pair is optimal. In this context, we pose the following questions:

Remark 1. *For the (3, 3) canonical cache network are there caching schemes achieving the memory rate pairs $(2/3, 4/3)$ and $(7/6, 5/6)$? The existence of such schemes will lead to the complete characterization of the exact rate memory tradeoff.*

Remark 2. *Is it possible to extend the schemes in sections II and III to the case of a general (N, K) network? Can the lower bounds obtained be extended to match such a general scheme?*

APPENDIX

In this appendix we prove the identities (1) and (2). Consider the (N, K) canonical cache network, where each user k has a cache memory Z_k . Let $W_{[N]}$ represent the set $W_{[N]} = \{W_1, \dots, W_N\}$. Now consider a demand $\mathbf{d} = \{W_{d_1}, \dots, W_{d_K}\}$, where the user k request for a file W_{d_k} . In response to this demand the server broadcast a set of packets $X_{\mathbf{d}}$. We start with the identity (1). We have,

$$\begin{aligned} H(W_{d_k}, Z_k, X_{\mathbf{d}}) &\stackrel{(a)}{=} H(Z_k, X_{\mathbf{d}}) + H(W_{d_k} | Z_k, X_{\mathbf{d}}) \\ &\stackrel{(b)}{=} H(Z_k, X_{\mathbf{d}}) + 0 \\ &= H(Z_k, X_{\mathbf{d}}). \end{aligned}$$

Now consider the identity 2. We have,

$$\begin{aligned} H(W_{[N]}, Z_k, X_{\mathbf{d}}) &\stackrel{(a)}{=} H(W_{[N]}) + H(Z_k, X_{\mathbf{d}} | W_{[N]}) \\ &\stackrel{(c)}{=} H(W_{[N]}) + 0 \\ &= H(W_{[N]}), \end{aligned}$$

where (a) follows from entropy chain rule, (b) follows from the fact that each user k can recover the requested file from the received packets with the help of its cached contents and (c) follows from the fact that each user's cached contents and the packets broadcast in response to the demand \mathbf{d} are functions of files in the server.

Proof for lemma 4.

$$\begin{aligned}
4M+12R &\geq 4H(Z_1) + 12H(X_{\{A,B,C,D\}}) \\
&\geq H(Z_1, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}, X_{\{C,D,A,B\}}) + H(Z_2, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}, X_{\{C,A,D,B\}}) \\
&\quad + H(Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}) + H(Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(a)}{=} H(A, B, C, Z_1, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}, X_{\{C,D,A,B\}}) + H(A, B, C, Z_2, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}, X_{\{C,A,D,B\}}) \\
&\quad + H(A, B, C, Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}) + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(b)}{\geq} H(A, B, C, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}) + H(A, B, C, Z_1, Z_2, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}, X_{\{C,A,D,B\}}, X_{\{C,D,A,B\}}) \\
&\quad + H(A, B, C, Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}) + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(a)}{=} H(A, B, C, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}) + H(A, B, C, D, Z_1, Z_2, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}, X_{\{C,A,D,B\}}, X_{\{C,D,A,B\}}) \\
&\quad + H(A, B, C, Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}) + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(c)}{=} H(A, B, C, X_{\{A,B,C,D\}}, X_{\{B,C,D,A\}}) + H(A, B, C, D) + H(A, B, C, Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}) \\
&\quad + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(b)}{\geq} H(A, B, C, D) + H(A, B, C, X_{\{A,B,C,D\}}) + H(A, B, C, Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}, X_{\{B,C,D,A\}}) \\
&\quad + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(a)}{=} H(A, B, C, D) + H(A, B, C, X_{\{A,B,C,D\}}) + H(A, B, C, D, Z_3, X_{\{A,B,C,D\}}, X_{\{B,C,A,D\}}, X_{\{D,A,B,C\}}, X_{\{B,C,D,A\}}) \\
&\quad + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(c)}{=} 2H(A, B, C, D) + H(A, B, C, X_{\{A,B,C,D\}}) + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}) \\
&\stackrel{(b)}{\geq} 2H(A, B, C, D) + H(A, B, C) + H(A, B, C, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}, X_{\{A,B,C,D\}}) \\
&\stackrel{(a)}{=} 2H(A, B, C, D) + H(A, B, C) + H(A, B, C, D, Z_4, X_{\{D,A,B,C\}}, X_{\{C,D,A,B\}}, X_{\{B,C,D,A\}}, X_{\{A,B,C,D\}}) \\
&\stackrel{(c)}{=} 2H(A, B, C, D) + H(A, B, C) + H(A, B, C, D) \\
&= 2 \times 4 + 3 + 4 = 15,
\end{aligned}$$

where (a) follows from (1), (b) follows from submodularity of entropy and (c) follows from (2).

REFERENCES

- [1] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.
- [2] Z. Chen, P. Fan, and K. B. Letaief, “Fundamental limits of caching: Improved bounds for users with small buffers,” *IET Communications*, vol. 10, no. 17, pp. 2315–2318, 2016.
- [3] M. M. Amiri and D. Gündüz, “Fundamental limits of coded caching: Improved delivery rate-cache capacity tradeoff,” *IEEE Transactions on Communications*, vol. 65, no. 2, pp. 806–815, 2017.
- [4] J. Gómez-Vilardebó, “Fundamental limits of caching: Improved rate-memory trade-off with coded prefetching,” *IEEE Transactions on Communications*, 2018.
- [5] C. Tian and J. Chen, “Caching and delivery via interference elimination,” in *IEEE International Symposium on Information Theory*, 2016, pp. 830–834.
- [6] K. Wan, D. Tuninetti, and P. Piantanida, “On the optimality of uncoded cache placement,” in *IEEE Information Theory Workshop*, 2016, pp. 161–165.
- [7] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “The exact rate-memory tradeoff for caching with uncoded prefetching,” *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1281–1296, 2017.
- [8] H. Ghasemi and A. Ramamoorthy, “Improved lower bounds for coded caching,” *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4388–4413, 2017.
- [9] A. Sengupta, R. Tandon, and T. C. Clancy, “Improved approximation of storage-rate tradeoff for caching via new outer bounds,” in *IEEE International Symposium on Information Theory*, 2015, pp. 1691–1695.
- [10] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “Characterizing the rate-memory tradeoff in cache networks within a factor of 2,” in *IEEE International Symposium on Information Theory*, 2017, pp. 386–390.
- [11] M. A. Maddah-Ali and U. Niesen, “Decentralized coded caching attains order-optimal memory-rate tradeoff,” *IEEE/ACM Transactions on Networking*, vol. 23, no. 4, pp. 1029–1040, 2015.
- [12] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, “Hierarchical coded caching,” *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3212–3229, 2016.
- [13] U. Niesen and M. A. Maddah-Ali, “Coded caching with nonuniform demands,” *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 1146–1158, 2017.
- [14] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, “Multi-server coded caching,” *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7253–7271, 2016.
- [15] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, “Online coded caching,” *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, 2016.
- [16] C. Tian, “Symmetry, outer bounds, and code constructions: A computer-aided investigation on the fundamental limits of caching,” *MDPI Entropy*, vol. 20, no. 8, p. 603, 2018.
- [17] ———, “A note on the fundamental limits of coded caching,” *arXiv preprint arXiv:1503.00010*, 2015.

On the Optimality of Simple Han-Kobayashi Schemes for Gaussian Interference Channels

Ragini Chaluvadi, Srikrishna Bhashyam

Department of Electrical Engineering, Indian Institute of Technology Madras

Abstract—The Generalized Degrees of Freedom (GDoF) region of the 2-user Gaussian Interference Channel (GIC) was derived by Etkin, Tse and Wang. This GDoF region is achieved using the class of Han-Kobayashi (HK) schemes. For K -user GICs with $K > 2$, the GDoF region is not known completely. For the K -user GIC, Geng et al. derived the channel conditions under which Gaussian signalling and Treating Interference as Noise (TIN) is GDoF optimal. The TIN scheme also belongs to the class of HK schemes. In this paper, we derive conditions under which *Simple HK* (S-HK) schemes are GDoF optimal for general K -user GICs. Simple HK schemes are HK schemes with Gaussian signalling, no time sharing, and no private-common power splitting. The class of simple HK schemes includes the TIN scheme and schemes that involve various levels of interference decoding and cancellation at each receiver.

I. INTRODUCTION

The capacity region of the general K user Gaussian Interference Channel (GIC) is not known. Key capacity and sum capacity results for the 2-user GIC were obtained in [1–7]. The capacity region is known under strong interference conditions, i.e., when the interference can be decoded [1, 2]. The sum capacity is known under weak and mixed interference conditions [3–6]. The capacity region of the 2 user GIC within one bit and the Generalized Degrees of Freedom (GDoF) region was derived in [7]. The complete GDoF region in [7] was achieved by suitably chosen Han-Kobayashi (HK) schemes [8].

For the general K user GIC, the channel conditions under which treating interference as noise (TIN) is optimal in terms of generalized degrees of freedom (GDoF) were derived in [9]. In this paper, we generalize the GDoF optimality results for the TIN scheme in [9] to all *Simple HK* (S-HK) schemes. We derive the channel conditions under which each S-HK scheme is GDoF optimal for the K -user GIC. S-HK schemes are HK schemes with Gaussian signalling, no time sharing, and no private-common power splitting. S-HK schemes include the simple and practical TIN scheme and schemes

that involve various levels of interference decoding and cancellation at each receiver as special cases.

II. SYSTEM MODEL

The model of a K -user GIC is given by

$$y_k = \sum_{i=1}^K h_{ki} x_i + z_k, \quad \forall k \in [K] \triangleq \{1, 2, \dots, K\}, \quad (1)$$

where x_i is transmitted by transmitter i , h_{ij} is the channel coefficient from transmitter j to receiver i and $z_k \sim \mathcal{CN}(0, 1)$ is the additive white Gaussian noise at receiver k . Let P_i denote the transmit power constraint at transmitter i . We call Han-Kobayashi schemes with Gaussian signaling, no timesharing, and no common-private power splitting as *simple HK* (S-HK) schemes. Each S-HK scheme is specified by the sets $\{I(1), I(2), \dots, I(K)\}$, $I(i) \subseteq [K] \forall i$. At receiver i , interference from transmitters $j \in I(i)$ are treated as noise and interference from transmitters $j \in D(i) \triangleq \{[K] \setminus \{I(i), i\}\}$ are decoded. For the TIN scheme, $I(i) = [K] \setminus i \forall i$. As in [9], we define the following quantities:

$$\begin{aligned} \text{SNR}_i &\triangleq \max(1, |h_{ii}|^2 P_i), \\ \text{INR}_{ki} &\triangleq \max(1, |h_{ki}|^2 P_i), \quad i \neq k, \quad \forall i, k \\ \alpha_{ii} &\triangleq \frac{\log \text{SNR}_i}{\log P}, \quad \alpha_{ki} \triangleq \frac{\log \text{INR}_{ki}}{\log P}, \quad i \neq k, \forall i, k, \end{aligned}$$

where we take $P > 1$ as a nominal power value. α_{ki} is the channel strength from transmitter i to receiver k . Making the SNR or INR one when they are less than one does not affect GDoF or constant gap results. We can prove this by following the similar steps as done in [9, Appendix A]. Now, we can represent the original channel (1) as:

$$y_k = \sum_{i=1}^K \sqrt{P^{\alpha_{ki}}} e^{j\theta_{ki}} \tilde{x}_i + z_k, \quad \forall k \in [K]. \quad (2)$$

In this equivalent model, $\tilde{x}_i = x_i / \sqrt{P_i}$ is the transmit symbol of transmitter i , and the power constraint for each transmitter is $E[|\tilde{x}_i|^2] \leq 1, \forall i \in [K]$. We use the same definitions for the GDoF region as in [9].

III. GDOF OPTIMALITY CONDITIONS

In this section, we derive channel conditions under which S-HK schemes can achieve the whole GDoF region.

A. Achievable GDoF region of S-HK schemes

Let transmitter i use power $P^{r_i}, r_i \leq 0$. For a S-HK scheme specified by $\{I(1), I(2), \dots, I(K)\}$, the following rate is achievable at receiver i

$$R_i = \log(1 + \text{SINR}_i) = \log \left(1 + \frac{P^{\alpha_{ii} + r_i}}{1 + \sum_{j \in I(i)} P^{\alpha_{ij} + r_j}} \right),$$

assuming that each receiver i can decode interference from transmitters in $D(i)$. The corresponding GDoF achieved by user i is

$$d_i = \max\{0, \alpha_{ii} + r_i - \max\{0, \max_{j \in I(i)} (\alpha_{ij} + r_j)\}\}. \quad (3)$$

The achievable GDoF region denoted by \mathcal{P}^* is the set of all K tuples (d_1, d_2, \dots, d_K) for which there exist r_i 's, $r_i \leq 0$, $i \in [K]$, such that (3) holds for all $i \in [K]$.

Sufficient conditions under which each receiver i can decode interference from transmitters in $D(i)$ are derived here.

Lemma 1. *In the K -user GIC with Gaussian signaling, if at each receiver j , we arrange α_{ij} 's for $i \in [K] \setminus \{j\}$ in descending order and denote the user corresponding to r^{th} position in the sequence as j_r . Then, the interference from the first k_j transmitters in $[K] \setminus \{j\}$ can be successively decoded, if*

$$\alpha_{jj_r} > \alpha_{j_r j_r} + \max\{\alpha_{jj}, \max_{s > r} \{\alpha_{j_s s}\}\} \quad \forall r \in \{1, \dots, k_j\}. \quad (4)$$

Proof. Under Gaussian signaling, interference from transmitter i to receiver j is decodable if

$$I(x_i; y_i | x_r, r \in D(i)) < I(x_i; y_j), \quad (5)$$

or, equivalently,

$$\begin{aligned} \left(\frac{\sum_{k \in I(i)} |h_{ik}|^2 P_k + 1}{|h_{ii}|^2} \right) &> \left(\frac{\sum_{k \neq i} |h_{jk}|^2 P_k + 1}{|h_{ji}|^2} \right), \\ \implies P^{\alpha_{ji}} &> P^{\alpha_{ii}} \left(\frac{\sum_{k \neq i} P^{\alpha_{jk}} + 1}{\sum_{k \in I(i)} P^{\alpha_{ik}} + 1} \right). \end{aligned}$$

Since, $\alpha_{ij} \geq 0, \forall i, j$, the above condition is satisfied if

$$P^{\alpha_{ji}} > P^{\alpha_{ii}} \cdot \left(\frac{K}{1 + |I(i)|} \cdot \max_{k \neq i} P^{\alpha_{jk}} \right).$$

Taking $\log_P(\cdot)$ and letting $P \rightarrow \infty$, we get the condition

$$\alpha_{ji} > \alpha_{ii} + \max_{k \neq i} \alpha_{jk}.$$

Now, if we arrange α_{ij} 's in descending order, then the interference from the first k_j transmitters in $[K] \setminus \{j\}$ can be successively decoded if (4) is satisfied. \square

B. GDoF Optimality of S-HK schemes

The following theorem gives channel conditions under which a S-HK scheme defined by $\{I(i), \forall i \in [K]\}$ is GDoF optimal. The corresponding GDoF region is also specified.

Theorem 1. *For a K -user GIC, at each receiver $j \in [K]$ if we arrange the indices of transmitters $i \in [K] \setminus \{j\}$ in the descending order of channel strengths α_{ji} and let j_r be the r^{th} index in the sequence. Then, if the channel strengths satisfy*

$$\alpha_{jj_r} > \alpha_{j_r j_r} + \max\{\alpha_{jj}, \max_{s > r} \{\alpha_{j_s s}\}\}, \quad \forall r \in \{1, \dots, k_j\}, \quad (6)$$

and

$$\alpha_{ll} \geq \max_{k \in I(l)} \alpha_{lk} + \max_{m: i \in I(m)} \alpha_{ml}, \quad \forall l \in [K]. \quad (7)$$

where $I(l) = \{k_l + 1, \dots, K - 1\}, \forall l \in [K]$, then the S-HK scheme defined by $\{I(l), \forall l \in [K]\}$ is GDoF optimal. And the corresponding GDoF region is the set of (d_1, d_2, \dots, d_K) satisfying

$$\begin{aligned} (a) \quad & 0 \leq d_i \leq \alpha_{ii}, \forall i \\ (b) \quad & \sum_{j=i}^m d_{i_j} \leq \sum_{j=1}^m [\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}], \quad \forall m \in \{2, 3, \dots, K\}, \\ & \forall (i_1, i_2, \dots, i_m) \in \pi_K \text{ s.t. } i_{k+1} \in I(i_k), \\ & \forall k = 0, \dots, m-1, i_0 = i_m, \\ (c) \quad & \sum_{j=i}^m d_{i_j} \leq \alpha_{i_1 i_1} + \sum_{j=2}^m [\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}], \\ & \forall m \in \{2, 3, \dots, K\}, \\ & \forall (i_1, i_2, \dots, i_m), \text{s.t. } i_{k+1} \in I(i_k), \\ & \forall k = 1, \dots, m-1, i_1 \notin I(i_m), \end{aligned} \quad (8)$$

where π_K is the set of all possible cyclic sequences of all subsets of $[K]$ with cardinality no less than 2.

Proof. As in [9], we consider a polyhedral S-HK scheme by ignoring the first $\max\{0, \dots\}$ term in (3), and obtain the polyhedral S-HK GDoF region $\mathcal{P} \subseteq \mathcal{P}^*$. GDoF region of the polyhedral S-HK \mathcal{P} is the set of tuples

(d_1, d_2, \dots, d_K) for which there exists r_i 's, $i \in [K]$, such that

$$r_i \leq 0, \forall i \in [K] \quad (9)$$

$$d_i \geq 0, \forall i \in [K] \quad (10)$$

$$d_i = \alpha_{ii} + r_i - \max\{0, \max_{j \in I(i)} (\alpha_{ij} + r_j)\}, \forall i \in [K] \quad (11)$$

Removing the first $\max\{0, \dots\}$ term in (3) can make d_i 's negative which is not valid. So, (10) is imposed. Inequalities (9)-(11) are equivalent to the set of inequalities (12)-(15) (proof is similar to that in [9]).

$$r_i \leq 0, \forall i \in [K] \quad (12)$$

$$d_i \geq 0, \forall i \in [K] \quad (13)$$

$$d_i \leq \alpha_{ii} + r_i, \forall i \in [K] \quad (14)$$

$$d_i \leq \alpha_{ii} + r_i - (\alpha_{ij} + r_j), \forall i \in [K], j \in I(i) \quad (15)$$

The region \mathcal{P} can be characterized as follows. Define a directed graph $D = (V, A)$ where

$$V = \{v_1, \dots, v_K, u\}$$

$$A = A_1 \cup A_2 \cup A_3$$

$$A_1 = \{(v_i, v_j) : i \in [K], j \in I(i)\}$$

$$A_2 = \{(v_i, u) : i \in [K]\}$$

$$A_3 = \{(u, v_i) : i \in [K]\}$$

Assign lengths $l(a)$ to every arc $a \in A$ as follows

$$l(v_i, v_j) = \alpha_{ii} - d_i - \alpha_{ij}, \forall i \in [K], j \in I(i)$$

$$l(v_i, u) = \alpha_{ii} - d_i$$

$$l(u, v_i) = 0.$$

Note that A_1 and $l(v_i, v_j)$ are different here compared to [9]. From [9, Lemma 1] and potential theorem [9], a GDoF tuple is in the region \mathcal{P} iff each directed circuit in the graph D has a non-negative length. All the circuits in D can be categorized in these three classes:

- Circuits of the form $(u, v_i, u), \forall i \in [K]$. From these circuits, we have the first set of conditions (a) in (8).
- Circuits of the form $(v_{i_0}, v_{i_1}, \dots, v_{i_m})$, where $i_0 = i_m, \forall (i_1, \dots, i_m) \in \pi_K$, such that $i_{k+1} \in I(i_k), \forall k \in [m-1]$, and $\forall m \in \{2, 3, \dots, K\}$. From these circuits, we have the second set of conditions (b) in (8).
- Circuits of the form $(u, v_{i_1}, \dots, v_{i_m}, u)$, for all $(i_1, \dots, i_m), \forall m \in \{2, 3, \dots, K\}$ such that $i_{k+1} \in I(i_k), \forall k \in \{1, \dots, m-1\}$. For these circuits we have

$$\sum_{j=1}^{m-1} [\alpha_{i_j i_j} - d_{i_j} - \alpha_{i_j i_{j+1}}] + [\alpha_{i_m i_m} - d_{i_m}] \geq 0 \quad (16)$$

If $i_1 \in I(i_m)$, (16) is redundant given conditions (b) in (8). If $i_1 \notin I(i_m)$, then we get the third set of conditions (c) in (8).

Different from the analysis for TIN in [9], we have 3 conditions above (instead of 2), and the additional requirement that $i_{k+1} \in I(i_k)$ in the last 2 conditions. The GDoF region \mathcal{P} can now be characterized, using only channel strengths, as the set of tuples (d_1, \dots, d_K) satisfying (8). This region is achievable under the conditions in (6).

It remains to be shown that the GDoF region is outer bounded by \mathcal{P} under the conditions in (7). It can be verified that the outer bound for GDoF region obtained from Theorem 2 below, by setting $G(i) = I(i)$, matches \mathcal{P} under condition (7). \square

Theorem 2. For the K -user IC described by (1), let $G(i) \subseteq [K], \forall i \in [K]$. The capacity region is included in the set of rate tuples (R_1, R_2, \dots, R_K) such that

$$R_i \leq \log(1 + |h_{ii}|^2 P_i), \forall i \in [K] \quad (17)$$

$$\sum_{j=1}^m R_{i_j} \leq \sum_{j=1}^m \log \left(1 + |h_{i_j i_{j+1}}|^2 P_{i_{j+1}} + \frac{|h_{i_j i_j}|^2 P_{i_j}}{1 + |h_{i_{j-1} i_j}|^2 P_{i_j}} \right),$$

$\forall m \in \{2, 3, \dots, K\}, \forall (i_1, i_2, \dots, i_m) \in \pi_K, \text{s.t}$

$i_{k+1} \in G(i_k), \forall k = 0, \dots, m-1, i_0 = i_m$

$$\begin{aligned} \sum_{j=1}^m R_{i_j} &\leq \sum_{j=1}^m \log \left(1 + |h_{i_j i_{j+1}}|^2 P_{i_{j+1}} + \frac{|h_{i_j i_j}|^2 P_{i_j}}{1 + |h_{i_{j-1} i_j}|^2 P_{i_j}} \right) \\ &\quad + \log(1 + |h_{i_m i_1}|^2 P_{i_1}), \\ &\forall m \in \{2, 3, \dots, K\}, \forall (i_1, i_2, \dots, i_m), \text{s.t} \\ &i_{k+1} \in G(i_k), \forall k = 1, \dots, m-1, i_1 \notin G(i_m) \end{aligned} \quad (19)$$

Proof. Bounds in (17) are simple. Bounds in (18) are already proved in [9, Thm. 3]. To prove (19), we modify (1) as follows. Consider (i_1, i_2, \dots, i_m) , such that $i_{k+1} \in G(i_k), \forall k = 1, \dots, m-1$.

- Eliminate all the users $i \in [K] \setminus \{i_1, i_2, \dots, i_m\}$ and their desired messages
- Remove all the interfering links but the links from transmitter i_j to receiver $i_{j-1}, \forall j \in \{1, 2, \dots, m\}$.

This modification does not hurt the rate of the users $i \in \{i_1, i_2, \dots, i_m\}$ in the original channel (1). Define

$$S_{i_j} = h_{i_{j-1} i_j} x_{i_j} + z_{i_{j-1}}, \forall j \in \{1, 2, \dots, m\}.$$

Note that the user indices are modulo- m .

To prove (19), consider (i_1, i_2, \dots, i_m) , such that $i_{k+1} \in G(i_k), \forall k = 1, \dots, m-1$, and $i_1 \notin G(i_m)$. Provide $S_{i_j}^n$ as a genie at each receiver $i_j, j \in \{1, 2, \dots, m-1\}$.

$1\}$ and provide $(S_{i_m}^n, x_{i_1}^n)$ at receiver i_m . For receiver i_m , we have (starting with Fano's inequality)

$$\begin{aligned} & n(R_{i_m} - \epsilon) \\ & \leq I(W_{i_m}; y_{i_m}^n, S_{i_m}^n, x_{i_1}^n) \\ & = h(y_{i_m}^n, S_{i_m}^n, x_{i_1}^n) - h(y_{i_m}^n, S_{i_m}^n, x_{i_1}^n | W_{i_m}) \\ & = h(S_{i_m}^n) + h(x_{i_1}^n | S_{i_m}^n) + h(y_{i_m}^n | S_{i_m}^n, x_{i_1}^n) \\ & \quad - h(S_{i_m}^n | W_{i_m}) - h(x_{i_1}^n | S_{i_m}^n, W_{i_m}) \\ & \quad - h(y_{i_m}^n | S_{i_m}^n, x_{i_1}^n, W_{i_m}) \\ & \stackrel{(a)}{\leq} h(S_{i_m}^n) + h(y_{i_m}^n | S_{i_m}^n) - h(z_{i_{m-1}}^n) \\ & \quad + [h(x_{i_1}^n | S_{i_m}^n) - h(x_{i_1}^n | S_{i_m}^n, W_{i_m})] - h(z_{i_m}^n) \\ & \stackrel{(b)}{=} h(S_{i_m}^n) - h(z_{i_m}^n) + h(y_{i_m}^n | S_{i_m}^n) - h(z_{i_{m-1}}^n) \end{aligned}$$

where (a) follows from the fact that conditioning reduces entropy, (b) follows since $\tilde{x}_{i_1}^n$ is independent of $W_{i_m}^n$. For receivers $i_j, j \neq m$, we have

$$\begin{aligned} & n(R_{i_j} - \epsilon) \\ & \leq h(S_{i_j}^n) + h(y_{i_j}^n | S_{i_j}^n) - h(S_{i_j}^n | W_{i_j}) - h(y_{i_j}^n | S_{i_j}^n, W_{i_j}) \\ & = h(S_{i_j}^n) + h(y_{i_j}^n | S_{i_j}^n) - h(z_{i_{j-1}}^n) - h(S_{i_{j+1}}^n) \end{aligned}$$

Adding above bounds, we get

$$\begin{aligned} & \sum_{j=1}^m R_{i_j} \\ & \leq \sum_{j=1}^m [h(y_{i_j}^n | S_{i_j}^n) - h(z_{i_j}^n)] + [h(s_{i_1}^n) - h(z_{i_m}^n)] \\ & \leq \sum_{j=1}^m \log \left(1 + |h_{i_j i_{j+1}}|^2 P_{i_{j+1}} + \frac{|h_{i_j i_j}|^2 P_{i_j}}{1 + |h_{i_{j-1} i_j}|^2 P_{i_j}} \right) \\ & \quad + \log(1 + |h_{i_m i_1}|^2 P_{i_j}). \end{aligned}$$

□

Corollary 1. For the K -user IC described by (2), with fixed $I(i) \subseteq [K], \forall i \in [K]$, when condition (7) is satisfied, its GDoF region is included in the set of GDoF tuples (d_1, d_2, \dots, d_K) such that (8) is satisfied.

Proof. This can be proved from Theorem 2, by choosing $G(i) = I(i)$.

From (17),

$$d_i = \lim_{P \rightarrow \infty} \frac{R_i}{\log P} \leq \lim_{P \rightarrow \infty} \frac{\log(1 + P^{\alpha_{ii}})}{\log P} = \alpha_{ii},$$

For all (i_1, \dots, i_m) , such that $i_{k+1} \in I(i_k), \forall k \in \{0, \dots, m-1\}$, we have

$$\sum_{j=1}^m d_{i_j} = \lim_{P \rightarrow \infty} \frac{\sum_{j=1}^m R_i}{\log P}$$

$$\begin{aligned} & \leq \lim_{P \rightarrow \infty} \frac{\sum_{j=1}^m \log \left(1 + P^{\alpha_{i_j i_{j+1}}} + \frac{P^{\alpha_{i_j i_j}}}{1 + P^{\alpha_{i_{j-1} i_j}}} \right)}{\log P} \\ & = \sum_{j=1}^m \max\{0, \alpha_{i_j i_{j+1}}, \alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}\} \\ & = \sum_{j=1}^m [\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}] \end{aligned}$$

where the last equality holds when (7) is satisfied.

For all (i_1, \dots, i_m) , such that $i_{k+1} \in I(i_k), \forall k = \{1, \dots, m-1\}, i_1 \notin I(i_m)$, we have

$$\begin{aligned} & \sum_{j=1}^m d_{i_j} \leq \lim_{P \rightarrow \infty} \frac{\log(1 + P^{\alpha_{i_m i_1}})}{\log P} + \lim_{P \rightarrow \infty} \\ & \quad \frac{\sum_{j=2}^m \log \left(1 + P^{\alpha_{i_j i_{j+1}}} + \frac{P^{\alpha_{i_j i_j}}}{1 + P^{\alpha_{i_{j-1} i_j}}} \right)}{\log P} \\ & \leq \alpha_{i_m i_1} + \sum_{j=1}^m \max\{0, \alpha_{i_j i_{j+1}}, \alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}\} \\ & = \alpha_{i_1 i_1} + \sum_{j=2}^m [\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}] \end{aligned}$$

where the last equality holds when (7) is satisfied. □

Theorem 3. (Constant Gap to Capacity) For the K -user GIC, under conditions (6), (7), the S-HK scheme defined by $\{I(i), \forall i \in [K]\}$ achieves rates within $\log_2(3 \cdot \max_{i \in [K]} (1 + |I(i)|))$ bits of the capacity region, where $|I(i)|$ is the cardinality of $I(i)$.

Proof. For proving this result, we use the similar steps as done in [9, Thm. 4]. For TIN scheme, $|I(i)| = K-1$ and we get a gap of $\log_2(3K)$ bits.

(Converse) From Theorem 2, we have

$$R_i \leq \log_2(1 + P^{\alpha_{ii}}) \leq \alpha_{ii} \log_2 P + 1 \quad (20)$$

For all $\forall (i_1, i_2, \dots, i_m) \in \pi_K$, s.t. $i_{k+1} \in I(i_k), \forall k = 0, \dots, m-1, i_0 = i_m$ and $m \in \{2, 3, \dots, K\}$ we have

$$\begin{aligned} & \sum_{j=1}^m R_{i_j} \\ & \leq \sum_{j=1}^m \log_2 \left(1 + P^{\alpha_{i_j i_{j+1}}} + \frac{P^{\alpha_{i_j i_j}}}{1 + P^{\alpha_{i_{j-1} i_j}}} \right) \\ & < \sum_{j=1}^m \log_2 \left(1 + P^{\alpha_{i_j i_{j+1}}} + \frac{P^{\alpha_{i_j i_j}}}{P^{\alpha_{i_{j-1} i_j}}} \right) \\ & = \sum_{j=1}^m \log_2 \left(\frac{P^{\alpha_{i_{j-1} i_j}} + P^{\alpha_{i_j i_{j+1}} + \alpha_{i_{j-1} i_j}} + P^{\alpha_{i_j i_j}}}{P^{\alpha_{i_{j-1} i_j}}} \right) \end{aligned}$$

$$\begin{aligned} &\stackrel{(a)}{\leq} \sum_{j=1}^m \log_2 \left(\frac{3P^{\alpha_{i_j i_j}}}{\alpha_{i_{j-1} i_j}} \right) \\ &= \sum_{j=1}^m [(\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P + \log_2 3]. \end{aligned} \quad (21)$$

where (a) is due to (7). Similary we have,

$$\begin{aligned} &\sum_{j=1}^m R_{i_j} \\ &\leq \alpha_{i_1 i_1} + \sum_{j=2}^m [(\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P + \log_2 3], \\ &\forall m \in \{2, 3, \dots, K\}, \forall (i_1, i_2, \dots, i_m), \text{s.t} \\ &i_{k+1} \in I(i_k), \forall k = 0, \dots, m-1, i_1 \notin I(i_m). \end{aligned} \quad (22)$$

(Achievability): We know that under (6),(7), if d'_i 's satisfy (8), then there exists $r_i \leq 0$ such that

$$r_i + \alpha_{ii} + r_i - \max\{0, \max_{j \in I(i)} (\alpha_{ij} + r_j)\} = d_i, \forall i \in [K] \quad (23)$$

Therefore the achievable S-HK rates R_i follows

$$\begin{aligned} R_i &= \log_2 \left(1 + \frac{P^{\alpha_{ii} + r_i}}{1 + \sum_{j \in I(i)} P^{\alpha_{ij} + r_j}} \right) \\ &\geq \log_2 \left(\frac{P^{\alpha_{ii} + r_i}}{P^0 + \sum_{j \in I(i)} P^{\alpha_{ij} + r_j}} \right) \\ &\stackrel{(b)}{\geq} \log_2 \left(\frac{P^{\alpha_{ii} + r_i}}{(1 + |I(i)|) P^{r_i + \alpha_{ii} - d_i}} \right) \\ &= d_i \log_2 P + \log_2 \left(\frac{1}{1 + |I(i)|} \right) \end{aligned} \quad (24)$$

where (b) follows from (23) Thus, we get the achievable region S-HK as the tuples (R_1, \dots, R_K) satisfying

$$R_i \leq \max \left\{ 0, \alpha_{ii} \log_2 P + \log \left(\frac{1}{1 + |I(i)|} \right) \right\} \quad (25)$$

$$\begin{aligned} \sum_{j=1}^m R_{i_j} &\leq \max \left\{ 0, \sum_{j=1}^m \left[(\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P \right. \right. \\ &\quad \left. \left. + \log_2 \left(\frac{1}{1 + |I(i_j)|} \right) \right] \right\}, \\ &\forall m \in \{2, 3, \dots, K\}, \forall (i_1, i_2, \dots, i_m) \in \pi_K, \text{s.t} \\ &i_{k+1} \in I(i_k), \forall k = 0, \dots, m-1, i_0 = i_m. \end{aligned} \quad (26)$$

$$\sum_{j=1}^m R_{i_j} \leq \max \left\{ 0, \alpha_{i_1 i_1} + \sum_{j=2}^m \left[(\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P \right. \right.$$

$$\begin{aligned} &\quad \left. \left. + \log_2 \left(\frac{1}{1 + |I(i_j)|} \right) \right] \right\}, \\ &\forall m \in \{2, 3, \dots, K\}, \forall (i_1, i_2, \dots, i_m), \text{s.t} \\ &i_{k+1} \in I(i_k), \forall k = 1, \dots, m-1, i_1 \notin I(i_m). \end{aligned} \quad (27)$$

Let $\sigma_1, \sigma_2, \sigma_3$ be the difference between the outer bounds (20)-(22) and their corresponding (25)-(27) achievable rates.

For bounding σ_1 , consider two cases

- $\alpha_{ii} \log_2 P + \log \left(\frac{1}{1 + |I(i)|} \right) \leq 0$: Here,
 $\sigma_1 = \alpha_{ii} \log_2 P + 1$
 $\stackrel{(c)}{\leq} 1 + \log_2(1 + |I(i)|) < \log_2(3(1 + |I(i)|))$
where (c) is due to our assumption.
- $\alpha_{ii} \log_2 P + \log \left(\frac{1}{1 + |I(i)|} \right) > 0$: Here, we have
 $\sigma_1 = 1 + \log_2(1 + |I(i)|) < \log_2(3(1 + |I(i)|))$

Therefore $\sigma_1 < \log_2(3(1 + |I(i)|))$.

Similarly for σ_2 , consider two cases

- $\sum_{j=1}^m (\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P + \log_2 \left(\frac{1}{1 + |I(i_j)|} \right) \leq 0$
Here, we have
 $\sigma_2 = \sum_{j=1}^m (\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P + \log_2 3$
 $\leq \sum_{j=1}^m \log_2(3(1 + |I(i_j)|))$
 $\leq m \log_2(3 \cdot \max_{i \in [K]} (1 + |I(i)|))$
- $\sum_{j=1}^m (\alpha_{i_j i_j} - \alpha_{i_{j-1} i_j}) \log_2 P + \log_2 \left(\frac{1}{1 + |I(i_j)|} \right) > 0$
Here,
 $\sigma_2 = \sum_{j=1}^m \log_2 3 + \log_2(1 + |I(i_j)|)$
 $\leq m \log_2(3 \cdot \max_{i \in [K]} (1 + |I(i)|))$

Therefore $\sigma_2 \leq m \log_2(3 \cdot \max_{i \in [K]} (1 + |I(i)|))$ and similarly we can prove $\sigma_3 \leq m \log_2(3 \cdot \max_{i \in [K]} (1 + |I(i)|))$.

Therefore, the S-HK scheme defined by $\{I(i), \forall i \in [K]\}$ achieves rates within $\log_2(3 \cdot \max_{i \in [K]} (1 + |I(i)|))$ bits of the capacity region. \square

In Appendix A, we explicitly write the GDOF region for the S-HK schemes for the 2-user GIC and verify that they agree with the results in [7]. We also provide the GDoF region for an example 3-user GIC.

IV. SUMMARY

We derived conditions under which simple HK schemes are GDoF optimal for the K -user Gaussian IC. This generalizes the GDoF result for the TIN scheme in [9] to all simple HK schemes. Thus, we now know the GDoF region for the K -user GIC for a larger set of channel conditions. The conditions for optimality are a combination of achievability conditions for interference decoding and converse conditions for treating part of the interference as noise.

REFERENCES

- [1] A. Carleial, "A case where interference does not reduce capacity (corresp.)," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 569–570, Sep 1975.
- [2] H. Sato, "The capacity of the gaussian interference channel under strong interference (corresp.)," *IEEE Transactions on Information Theory*, vol. 27, no. 6, pp. 786–788, Nov 1981.
- [3] V. S. Annapureddy and V. V. Veeravalli, "Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on the capacity region," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3032–3050, July 2009.
- [4] X. Shang, G. Kramer, and B. Chen, "A new outer bound and the noisy-interference sumrate capacity for gaussian interference channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 689–699, Feb 2009.
- [5] ———, "New outer bounds on the capacity region of gaussian interference channels," in *Proc. IEEE (ISIT), Toronto, ON, Canada*, Jul. 2008, pp. 245–249.
- [6] A. S. Motahari and A. K. Khandani, "Capacity bounds for the gaussian interference channel," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 620–643, Feb 2009.
- [7] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5534–5562, Dec 2008.
- [8] T. Han and K. Kobayashi, "A new achievable rate region for the interference channel," *IEEE Transactions on Information Theory*, vol. 27, no. 1, pp. 49–60, Jan 1981.
- [9] C. Geng, N. Naderializadeh, A. S. Avestimehr, and S. A. Jafar, "On the optimality of treating interference as noise," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 1753–1767, April 2015.

APPENDIX

A. GDoF of S-HK for a 2 user GIC

In [7], the GDoF region for a 2-user GIC was found by classifying all the 2-user channels into three groups, namely, weak interference channel ($\alpha_{12} < \alpha_{22}$ and $\alpha_{21} < \alpha_{11}$), mixed interference channel ($\alpha_{12} \geq \alpha_{22}$, $\alpha_{21} < \alpha_{11}$ or $\alpha_{12} < \alpha_{22}$, $\alpha_{21} \geq \alpha_{11}$) and strong interference channel ($\alpha_{12} \geq \alpha_{22}$, $\alpha_{21} \geq \alpha_{11}$).

For all the possible S-HK schemes of a 2 user GIC, the optimality conditions for these schemes and GDoF region are given in Table I. Channels satisfying the optimality conditions given in Table I for Schemes 1, 2, and 3 are weak interference, mixed interference, and strong interference channels, respectively, as defined in [7]. The GDoF region derived in [7] reduces to GDoF results in Table I under these optimality conditions.

S-HK scheme	Optimality conditions	GDoF region
1. $I(1) = \{2\}$, $I(2) = \{1\}$	$\alpha_{11} \geq \alpha_{12} + \alpha_{21}$, $\alpha_{22} \geq \alpha_{12} + \alpha_{21}$	$d_1 \leq \alpha_{11}$, $d_2 \leq \alpha_{22}$, $d_1 + d_2 \leq \alpha_{11} + \alpha_{22} - \alpha_{12} - \alpha_{21}$
2a. $I(1) = \{\}$, $I(2) = \{1\}$	$\alpha_{12} \geq \alpha_{11} + \alpha_{22}$, $\alpha_{11} \geq \alpha_{21}$, $\alpha_{22} \geq \alpha_{21}$	$d_1 \leq \alpha_{11}$, $d_2 \leq \alpha_{22}$, $d_1 + d_2 \leq \alpha_{11} + \alpha_{22} - \alpha_{21}$
2b. $I(1) = \{2\}$, $I(2) = \{\}$	$\alpha_{21} \geq \alpha_{11} + \alpha_{22}$, $\alpha_{11} \geq \alpha_{12}$, $\alpha_{22} \geq \alpha_{12}$	$d_1 \leq \alpha_{11}$, $d_2 \leq \alpha_{22}$, $d_1 + d_2 \leq \alpha_{11} + \alpha_{22} - \alpha_{12}$
3. $I(1) = \{\}$, $I(2) = \{\}$	$\alpha_{12} \geq \alpha_{11} + \alpha_{22}$, $\alpha_{21} \geq \alpha_{11} + \alpha_{22}$	$d_1 \leq \alpha_{11}$, $d_2 \leq \alpha_{22}$

TABLE I: Optimality conditions and GDoF region of S-HK schemes for a 2 user GIC.

Finally, we note that under the optimality conditions for S-HK scheme 1, i.e.,

$$\alpha_{11} \geq \alpha_{12} + \alpha_{21} \quad (28)$$

$$\alpha_{22} \geq \alpha_{12} + \alpha_{21}, \quad (29)$$

TIN is GDoF optimal. For a 2-user GIC, these conditions are also necessary for TIN to be GDoF optimal since the GDoF region of weak interference channels found in [7] reduces to GDoF of TIN scheme only when (28), (29) are satisfied.

B. 3-user GIC

Example 1. Consider a 3 user GIC with channel strengths.

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} = \begin{bmatrix} 1.5 & 3 & 0.5 \\ 0.5 & 1.5 & 3 \\ 3 & 1 & 1.5 \end{bmatrix}$$

For this channel, S-HK is GDoF optimal with $I(1) = \{3\}$, $I(2) = \{1\}$, $I(3) = \{2\}$ since the channel satisfies conditions (6), (7). The set of $(i_1, i_2, \dots, i_m) \in \pi_K$, such that $i_{k+1} \in I(i_k), \forall k = 1, \dots, m-1$ and $i_1 \in I(i_m)$ is empty, and the set (i_1, i_2, \dots, i_m) , such that $i_{k+1} \in I(i_k), \forall k = 1, \dots, m-1, i_1 \notin I(i_m)$ is $\{(1, 3), (2, 1), (3, 2), (1, 3, 2), (2, 1, 3), (3, 2, 1)\}$. The GDoF region is the set of (d_1, d_2, \dots, d_K) satisfying

$$0 \leq d_1, d_2, d_3 \leq 1.5$$

$$d_1 + d_3 \leq 2.5, \quad d_2 + d_1 \leq 2.5$$

$$d_3 + d_2 \leq 2$$

$$d_1 + d_3 + d_2 \leq 3$$

$$d_2 + d_1 + d_3 \leq 3.5$$

$$d_3 + d_2 + d_1 \leq 3.$$

A Minimax Theorem for Finite Blocklength Joint Source-Channel Coding over an AVC

Anuj Vora, Ankur Kulkarni
Systems and Control Engineering, IIT Bombay, Mumbai 400076
Email - anujvora@iitb.ac.in, kulkarni.ankur@iitb.ac.in

Abstract—We pose the finite blocklength communication problem in the presence of a jammer as a zero-sum game between the encoder-decoder team and the jammer, where the communicators, as well as the jammer, are allowed locally randomized strategies. The minimax value of the game corresponds to joint source-channel coding over an Arbitrarily Varying Channel (AVC), which in the channel coding setting is known to admit a strong converse. The communicating team’s problem is non-convex and hence, in general, a minimax theorem need not hold for this game. However, we show that an *approximate minimax* theorem holds in the sense that the minimax and maximin values of the game approach each other asymptotically. In particular, for rates above a critical threshold, both the minimax and maximin values approach unity. This result is stronger than the usual strong converse for channel coding over an AVC, which only says that the minimax value approaches unity for such rates.

I. INTRODUCTION

We consider a setting where a team of finite blocklength encoder and decoder attempt to communicate an i.i.d. source over a noisy channel whose state is controlled by a jammer. We formulate the above problem as a zero-sum game where the communicating team attempts to minimize the average probability of error while the jammer tries to maximize it. Finding the minimax or the upper value of the game amounts to a joint source-channel coding problem over an AVC. It is known that in the channel coding setting, the AVC admits a *strong converse*, which says that the capacity of the AVC is a sharp threshold for rates below which communication is possible with arbitrarily high probability while for rates above the capacity, it is impossible to communicate with non-zero probability. The maximin problem, on the other hand, provides a jammer’s perspective of the game where the objective is to maximize the smallest probability of error under all actions of the encoder-decoder team.

In this paper, we show that the above strong converse is a consequence of a deeper phenomenon: the zero-sum game posed above admits a near-saddle point whose value approaches zero for rates below the aforementioned threshold and unity for rates above the threshold. This result is stronger than the strong converse since it implies that the for rates above the threshold the upper *and* lower values of the game tend to unity, whereas the usual strong converse talks only about the upper value. Moreover, our results extend to joint source-channel coding over an AVC, wherein the corresponding threshold is given by the ratio of the capacity of the AVC to the entropy of the source.

For a fixed action of the jammer, the optimization problem for the communicating team is non-convex due to the non-classical information structure [1] and hence a saddle point need

not exist for the zero-sum game. Recently, the authors in [2] showed that although the problem is non-convex, it nevertheless possesses a hidden convexity. Specifically, they demonstrate that for large blocklengths, the problem can be approximated arbitrarily closely by a linear programming based relaxation. Consequently, one may suppose that an *approximate minimax* theorem may hold for the zero-sum game. Our results validate this intuition and the difference between the upper and lower values of the game becomes arbitrarily small as the blocklength goes to infinity.

A closely related but distinct setting has been studied in [3]. There the authors consider only the channel coding problem and the action of the jammer is fixed throughout the transmission, making the upper value problem equivalent to coding for the *compound channel*.

Our proof proceeds by deriving an upper and lower bound for the upper and the lower values of the game. The upper bound is derived by constructing an achievability scheme using separate source-channel coding. The lower bound follows from the linear programming relaxation method from [2].

This paper is organized as follows. We formulate the problem in Section III. The lower bound is derived in Section IV and the upper bound is derived in Section V. The corresponding asymptotic analysis is done in Section VI and Section VII concludes the paper.

II. NOTATION AND PRELIMINARIES

All random variables are defined on an underlying probability space with measure \mathbb{P} . Random variables are represented with uppercase letters X and their instances are denoted by lower case letters x ; unless otherwise stated these are vectors whose length will be understood from the context. Blackboard letters \mathbb{X}, \mathbb{Y} etc. are used to represent the corresponding single-letter random variable. Calligraphic letters \mathcal{X}, \mathcal{Y} etc. denote spaces of single-letter random variables. $\mathcal{P}(\mathcal{X})$ denotes the set of all probability distributions on a space \mathcal{X} and a particular distribution is represented as $P_{\mathbb{X}}$. For any distributions P_X and $P_{Y|X}$, we define $(P_X \times P_{Y|X})(x, y) = P_X(x)P_{Y|X}(y|x)$ and $(P_X P_{Y|X})(y) = \sum_x P_X(x)P_{Y|X}(y|x)$. The type of a sequence $x \in \mathcal{X}^n$ is the empirical distribution $P_x \in \mathcal{P}(\mathcal{X})$, given by $P_x(\bullet) \equiv \frac{|\{i: x_i = \bullet\}|}{n}$. The joint type of x, y is denoted as $P_{x,y}$. The set $\mathcal{P}_n(\mathcal{X}) \subseteq \mathcal{P}(\mathcal{X})$ denotes the set of all types of sequence in \mathcal{X}^n . $T(P)$ denotes the set of sequences with type P .

III. PROBLEM FORMULATION

Consider a family of channels $\mathcal{V} := \{P_{\mathbb{Y}|\mathbb{X}, \Theta}(y|x, \theta), x \in \mathcal{X}, y \in \mathcal{Y}, \theta \in \mathcal{T}\}$, having common input spaces and output spaces denoted by the finite sets \mathcal{X} and \mathcal{Y} , respectively, where

each channel is indexed by the parameter θ , called the state of the channel, drawn from a finite space \mathcal{T} . Let \mathcal{S} be a finite space. Suppose a random source message $S \in \mathcal{S}^k, k \in \mathbb{N}$, generated i.i.d. according a fixed distribution $P_{\mathbb{S}} \in \mathcal{P}(\mathcal{S})$, is to be communicated over this family of channels where a jammer can choose a channel from the set \mathcal{V} for every transmission. An encoder encodes the message S into the channel input string $X \in \mathcal{X}^n, n \in \mathbb{N}$ according to a law $Q_{X|S} \in \mathcal{P}(\mathcal{X}^n|\mathcal{S}^k)$. The channel output string $Y \in \mathcal{Y}^n$ is decoded by the decoder to $\hat{S} \in \mathcal{S}^k$ according to a law $Q_{\hat{S}|Y} \in \mathcal{P}(\mathcal{S}^k|\mathcal{Y}^n)$ and an error is said to occur if $\hat{S} \neq S$. The jammer selects the channels by choosing a random state sequence $\Theta \in \mathcal{T}^n$ distributed according to $q \in \mathcal{P}(\mathcal{T}^n)$; $\Theta \in \mathcal{T}$ denotes the single-letterized random variable. We assume that the encoder and decoder do not know the actions of the jammer and that the jammer also does not have any information about the actions of the encoder and decoder or the source message.

We assume the channel behaviour is memoryless. Thus, the resulting channel can be modeled as the following discrete memoryless AVC defined by choosing a channel from the family \mathcal{V} for every transmission and constructing

$$P_{Y|X,\Theta}(y|x, \theta) = \prod_{i=1}^n P_{Y|\mathbb{X},\Theta}(y_i|x_i, \theta_i), \quad (1)$$

where (1) governs the probability of receiving the output sequence $y = (y_1, \dots, y_n)$ when the input sequence is $x = (x_1, \dots, x_n)$ and the state sequence is $\theta = (\theta_1, \dots, \theta_n)$. The rate of communication in this setting is defined as $R = \frac{k}{n}$.

The probability of error is given as

$$\begin{aligned} \mathbb{P}(\hat{S} \neq S) &= \sum_{s,x,y,\hat{s},\theta} \mathbb{I}\{\hat{s} \neq s\} q(\theta) P_S(s) Q_{X|S}(x|s) \\ &\quad \times P_{Y|X,\Theta}(y|x, \theta) Q_{\hat{S}|Y}(\hat{s}|y). \end{aligned} \quad (2)$$

We assume that the encoder and decoder aim to minimize the probability of error by choosing stochastic codes $(Q_{X|S}, Q_{\hat{S}|Y})$ while the jammer tries to maximize it by choosing the distribution q . Thus, for every pair of (k, n) , the problem is a zero-sum game between the encoder-decoder team and the jammer with the probability of error as the cost function.

The minimax or upper value of the game is given by

$$\begin{aligned} \bar{v}(k, n) &= \min_{Q_{X|S}, Q_{\hat{S}|Y}} \max_q \mathbb{P}(\hat{S} \neq S) \\ \text{s.t. } &Q_{X|S} \in \mathcal{P}(\mathcal{X}^n|\mathcal{S}^k), Q_{\hat{S}|Y} \in \mathcal{P}(\mathcal{S}^k|\mathcal{Y}^n), q \in \mathcal{P}(\mathcal{T}^n), \end{aligned}$$

and the maximin or the lower value of the game is given by

$$\begin{aligned} \underline{v}(k, n) &= \max_q \min_{Q_{X|S}, Q_{\hat{S}|Y}} \mathbb{P}(\hat{S} \neq S) \\ \text{s.t. } &Q_{X|S} \in \mathcal{P}(\mathcal{X}^n|\mathcal{S}^k), Q_{\hat{S}|Y} \in \mathcal{P}(\mathcal{S}^k|\mathcal{Y}^n), q \in \mathcal{P}(\mathcal{T}^n). \end{aligned}$$

Clearly, we have that $\bar{v}(k, n) \geq \underline{v}(k, n)$.

It can be observed that the minimax problem of the zero-sum game is a joint source-channel coding problem over an AVC with stochastic codes, since the encoder and decoder search for stochastic codes which minimize the worst case probability of error and it is optimal for the jammer to pick a deterministic sequence of states. For an AVC, a necessary and sufficient condition for the stochastic code capacity to be positive is *non-symmetrizability* [4]. An AVC is said to be non-symmetrizable

if there does not exist any distribution $P_{\Theta|\mathbb{X}} \in \mathcal{P}(\mathcal{T}|\mathcal{X})$ such that $\forall x \in \mathcal{X}, x' \in \mathcal{X}, y \in \mathcal{Y}$, we have

$$\sum_{\theta \in \mathcal{T}} P_{\Theta|\mathbb{X}}(\theta|x) P_{\mathbb{Y}|\mathbb{X},\Theta}(y|x', \theta) = \sum_{\theta \in \mathcal{T}} P_{\Theta|\mathbb{X}}(\theta|x') P_{\mathbb{Y}|\mathbb{X},\Theta}(y|x, \theta).$$

In this paper, it is assumed that the AVC under study is non-symmetrizable. For a non-symmetrizable AVC, the stochastic code capacity for maximum (and average) probability of error criterion is given from [5] as

$$C = \max_{P_{\mathbb{X}} \in \mathcal{P}(\mathcal{X})} \min_{q_{\Theta} \in \mathcal{P}(\mathcal{T})} I(\mathbb{X}; \mathbb{Y}_q), \quad (3)$$

where $I(\mathbb{X}; \mathbb{Y}_q)$ is the mutual information between \mathbb{X} and \mathbb{Y}_q , with $\mathbb{Y}_q \sim (q_{\Theta} P_{\mathbb{Y}|\mathbb{X},\Theta})(y|x) := \sum_{\theta \in \mathcal{T}} q_{\Theta}(\theta) P_{\mathbb{Y}|\mathbb{X},\Theta}(y|x, \theta)$ where $x \in \mathcal{X}, y \in \mathcal{Y}$. Also, a strong converse holds for the channel coding over an AVC (Corr. 12.3 [6]).

In a joint source-channel coding problem over a DMC without a jammer, asymptotically vanishing probability of error can be achieved for rates below $\frac{C'}{H(\mathbb{S})}$, where C' is the capacity of the DMC and $H(\mathbb{S})$ is the entropy of the source. Further, the probability of error goes to one for rates above $\frac{C'}{H(\mathbb{S})}$ [7]. In this paper, we show that a similar but a stronger results holds, where $\bar{v}(k, n)$ and $\underline{v}(k, n)$ approach each other as $k, n \rightarrow \infty$ at a fixed rate $\frac{k}{n}$. In particular, the values tend to zero for $\frac{k}{n} < \frac{C}{H(\mathbb{S})}$ and they tend to unity for $\frac{k}{n} > \frac{C}{H(\mathbb{S})}$, where C is the capacity of the AVC. As a corollary, we get that the strong converse holds for the joint source-channel coding over an AVC.

IV. LOWER BOUND ON THE MAXIMIN VALUE

In this section, we derive a lower bound on $\underline{v}(k, n)$, by relaxing the inner minimization over $(Q_{X|S}, Q_{\hat{S}|Y})$ in the maximin problem. For each $q \in \mathcal{P}(\mathcal{T}^n)$, the minimization can be written as

$$\begin{aligned} \text{SC}(q) &\min_{Q_{X|S}, Q_{\hat{S}|Y}} \sum_{s,x,y,\hat{s}} \mathbb{I}\{\hat{s} \neq s\} Q(s, x, y, \hat{s}) \\ &\quad Q(s, x, y, \hat{s}) \equiv P_S(s) Q_{X|S}(x|s) P_{Y_q|X}(y|x) \\ &\quad \times Q_{\hat{S}|Y}(\hat{s}|y), \\ \text{s.t. } &\sum_x Q_{X|S}(x|s) = 1 \quad \forall s, \\ &\sum_{\hat{s}} Q_{\hat{S}|Y}(\hat{s}|y) = 1 \quad \forall y, \\ &Q_{X|S}(x|s), Q_{\hat{S}|Y}(\hat{s}|y) \geq 0 \quad \forall s, x, y, \hat{s}, \end{aligned}$$

where $P_{Y_q|X}(y|x) := \sum_{\theta \in \mathcal{T}^n} q(\theta) P_{Y|\mathbb{X},\Theta}(y|x, \theta)$.

The above problem is shown to be non-convex in the space of the distributions $(Q_{X|S}, Q_{\hat{S}|Y})$ [1]. A particular line of approach for such problems is to derive a convex relaxation by containing the non-convex feasible region within a convex set. We consider a linear programming relaxation presented in [2] derived by a lift-and-project like method.

$$\begin{aligned} \text{LP}(q) &\min_{Q_{X|S}, Q_{\hat{S}|Y}, V} \sum_{s,x,y,\hat{s}} \mathbb{I}\{s \neq \hat{s}\} P_S(s) P_{Y_q|X}(y|x) V(s, x, y, \hat{s}) \\ &\quad \sum_x Q_{X|S}(x|s) = 1 : \gamma_q^a(s) \quad \forall s, \\ &\quad \sum_{\hat{s}} Q_{\hat{S}|Y}(\hat{s}|y) = 1 : \gamma_q^b(y) \quad \forall y, \\ \text{s.t. } &\sum_x V(s, x, y, \hat{s}) - Q_{\hat{S}|Y}(\hat{s}|y) = 0 : \lambda_q^a(s, y, \hat{s}) \quad \forall s, y, \hat{s}, \\ &\sum_{\hat{s}} V(s, x, y, \hat{s}) - Q_{X|S}(x|s) = 0 : \lambda_q^b(s, x, y) \quad \forall s, x, y, \\ &Q_{X|S}(x|s), Q_{\hat{S}|Y}(\hat{s}|y) \geq 0 \quad \forall s, x, y, \hat{s}, \\ &V(s, x, y, \hat{s}) \geq 0 \quad \forall s, x, y, \hat{s}. \end{aligned}$$

Clearly, we have $\text{OPT}(\text{SC}(q)) \geq \text{OPT}(\text{LP}(q)) \forall q \in \mathcal{P}(\mathcal{T}^n)$. The corresponding dual program is given as follows.

$$\begin{aligned} \text{DP}(q) &= \max_{\gamma_q^a, \gamma_q^b, \lambda_q^a, \lambda_q^b} \sum_s \gamma_q^a(s) + \sum_y \gamma_q^b(y) \\ \text{s.t.} &\quad \gamma_q^a(s) - \sum_y \lambda_q^b(s, x, y) \leq 0 \quad \forall s, x, \\ &\quad \gamma_q^b(y) - \sum_s \lambda_q^a(s, y, \hat{s}) \leq 0 \quad \forall y, \hat{s}, \\ &\quad \lambda_q^a(s, y, \hat{s}) + \lambda_q^b(s, x, y) \leq \Pi(s, x, y, \hat{s}) \quad \forall s, x, y, \hat{s}, \end{aligned}$$

where $\Pi(s, x, y, \hat{s}) \equiv \mathbb{I}\{\hat{s} \neq s\} P_S(s) P_{Y_q|X}(y|x)$ and the functions $\gamma^a : \mathcal{S}^k \rightarrow \mathbb{R}$, $\gamma^b : \mathcal{Y}^n \rightarrow \mathbb{R}$, $\lambda^a : \mathcal{S}^k \times \mathcal{Y}^n \times \mathcal{S}^k \rightarrow \mathbb{R}$ and $\lambda^b : \mathcal{S}^k \times \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \mathbb{R}$ are Lagrange multipliers. From LP duality it follows that $\text{OPT}(\text{SC}(q)) \geq \text{OPT}(\text{LP}(q)) = \text{OPT}(\text{DP}(q)) \forall q$ and from [3] we get

$$\begin{aligned} \bar{\nu}(k, n) &\geq \underline{\nu}(k, n) = \max_{q \in \mathcal{P}(\mathcal{T}^n)} \text{OPT}(\text{SC}(q)) \\ &\geq \max_{q \in \mathcal{P}(\mathcal{T}^n)} \text{OPT}(\text{LP}(q)) = \max_{q \in \mathcal{P}(\mathcal{T}^n)} \text{OPT}(\text{DP}(q)). \end{aligned} \quad (4)$$

Since the optimal value $\max_q \text{OPT}(\text{DP}(q))$ is a lower bound for the minimax, as well as, the maximin value of the zero-sum game, it is sufficient to derive a feasible solution of the $\text{DP}(q)$ to compute a lower bound for the values of the game.

Lemma 4.1: The following functions are feasible for $\text{DP}(q)$

$$\begin{aligned} \lambda_q^a(s, y, \hat{s}) &= -\mathbb{I}\{\hat{s} = s\} \exp(-\gamma) P_{\bar{Y}_q}(y), \\ \lambda_q^b(s, x, y) &= P_S(s) \min \left\{ P_{Y_q|X}(y|x), \frac{P_{\bar{Y}_q}(y)}{P_S(s)} \exp(-\gamma) \right\}, \\ \gamma_q^a(s) &= \min_x \sum_y \lambda_q^b(s, x, y), \quad \gamma_q^b(y) = \min_{\hat{s}} \sum_s \lambda_q^a(s, y, \hat{s}), \end{aligned}$$

where $\gamma > 0$ and $P_{\bar{Y}_q}(y) = \sum_{\theta \in \mathcal{T}^n} q(\theta) P_{\bar{Y}|\Theta}(y|\theta)$ and $P_{\bar{Y}|\Theta}$ is any distribution in $\mathcal{P}(\mathcal{Y}^n | \mathcal{T}^n)$.

Proof : For $P_S(s) P_{Y_q|X}(y|x) \leq P_{\bar{Y}_q}(y) \exp(-\gamma)$, We have that $\lambda_q^a(s, y, \hat{s}) + \lambda_q^b(s, x, y) \leq -\mathbb{I}\{\hat{s} = s\} \exp(-\gamma) P_{\bar{Y}_q}(y) + P_S(s) P_{Y_q|X}(y|x) \leq \mathbb{I}\{\hat{s} \neq s\} P_S(s) P_{Y_q|X}(y|x)$. The other case $P_S(s) P_{Y_q|X}(y|x) > P_{\bar{Y}_q}(y) \exp(-\gamma)$ follows similarly. Thus, $\lambda_q^a(s, y, \hat{s})$ and $\lambda_q^b(s, x, y)$ are feasible for $\text{DP}(q)$. Further, $\gamma_q^a(s)$ and $\gamma_q^b(y)$ are feasible by construction. ■

We now compute the dual cost of $\text{DP}(q)$ for the above feasible solution. Below, the probability of an even A under the measure induced by a distribution P is denoted by $P\{A\} := \sum_x \mathbb{I}\{x \in A\} P(x)$.

Theorem 4.1: The value $\underline{\nu}(k, n)$ is lower bounded as

$$\begin{aligned} \underline{\nu}(k, n) &\geq \max_q \text{OPT}(\text{DP}(q)) \geq \max_{q, P_{\bar{Y}_q}} \sup_{\gamma > 0} \sum_s P_S(s) \\ &\times \min_{x \in \mathcal{X}^n} \left[P_{Y_q|X=x} \left\{ \frac{P_{Y_q|X}(Y|x)}{P_{\bar{Y}_q}(Y)} \leq \frac{\exp(-\gamma)}{P_S(s)} \right\} \right. \\ &\left. + \frac{\exp(-\gamma)}{P_S(s)} \left(P_{\bar{Y}_q} \left\{ \frac{P_{Y_q|X}(Y|x)}{P_{\bar{Y}_q}(Y)} > \frac{\exp(-\gamma)}{P_S(s)} \right\} - 1 \right) \right]. \end{aligned} \quad (5)$$

Proof : From $\text{DP}(q)$, we have that $\text{OPT}(\text{DP}(q)) \geq \sum_s \min_x \sum_y \lambda_q^b(s, x, y) + \sum_y \min_{\hat{s}} \sum_s \lambda_q^a(s, y, \hat{s})$. Taking $\lambda_q^b(s, x, y)$ and $\lambda_q^a(s, y, \hat{s})$ as in Lemma 4.1 and taking maximum over γ , $P_{\bar{Y}_q}$ and q we get the expression on the RHS of (5). The required bound follows from equation (4). ■

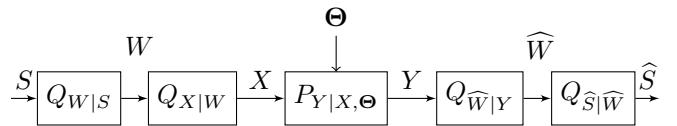


Fig. 1. Separate Source-Channel Coding

V. UPPER BOUND ON THE MINIMAX VALUE

To construct an upper bound on the minimax value $\bar{\nu}(k, n)$, we construct a separate source code and a channel code independent of each other. The source-channel code is then assembled by taking the composition of the source code and the channel code. The separate source-channel coding setup is given in Figure 1. Let the output random variable of the source encoder be $W \in \mathcal{W} := \{1, \dots, M\}$, where $M \in \mathbb{N}$, and the input random variable of the source decoder be $\widehat{W} \in \widehat{\mathcal{W}}$. The source code is defined as the pair of conditional distributions $Q_{W|S} \in \mathcal{P}(\mathcal{W}|\mathcal{S}^k)$, $Q_{\widehat{S}|\widehat{W}} \in \mathcal{P}(\mathcal{S}^k|\mathcal{W})$ given as $Q_{W|S}(w|s) = \mathbb{I}\{f_S(s) = w\}$, $Q_{\widehat{S}|\widehat{W}}(\widehat{s}|\widehat{w}) = \mathbb{I}\{\varphi_S(\widehat{w}) = \widehat{s}\}$, where (f_S, φ_S) are deterministic functions defined as $f_S : \mathcal{S}^k \rightarrow \mathcal{W}$, $\varphi_S : \mathcal{W} \rightarrow \mathcal{S}^k$. The channel code is defined as the pair of conditional distributions $Q_{X|W} \in \mathcal{P}(\mathcal{X}^n|\mathcal{W})$, $Q_{\widehat{W}|Y} \in \mathcal{P}(\mathcal{W}|\mathcal{Y}^n)$, where $Q_{X|W}$ is a stochastic encoder and $Q_{\widehat{W}|Y}(\widehat{w}|y) = \mathbb{I}\{\varphi_C(y) = \widehat{w}\}$, where φ_C is a deterministic function defined as $\varphi_C : \mathcal{Y}^n \rightarrow \mathcal{W}$. The composition of the two codes gives the following separate source-channel code

$$\begin{aligned} Q_{X|S}(x|s) &:= \sum_w \mathbb{I}\{f_S(s) = w\} Q_{X|W}(x|w) \\ Q_{\widehat{S}|Y}(\widehat{s}|y) &:= \sum_{\widehat{w}} \mathbb{I}\{\varphi_C(y) = \widehat{w}\} \mathbb{I}\{\varphi_S(\widehat{w}) = \widehat{s}\}. \end{aligned} \quad (6)$$

From the above definition, it suffices to construct the functions (f_S, φ_S) , the stochastic encoder $Q_{X|W}$ and the function φ_C to construct a stochastic source-channel code.

A. Channel code

In this section, we consider two kinds of channel codes, a deterministic code for average probability of error and a random code for the maximum probability of error. Using these codes we then construct a channel code with stochastic encoder and deterministic decoder. For a code (f, φ) and a distribution ψ on the set of channel codes $\{(f, \varphi) | f : \mathcal{W} \rightarrow \mathcal{X}^n, \varphi : \mathcal{Y}^n \rightarrow \mathcal{W}\}$, we define the following,

$$\begin{aligned} e_{m, \theta}(f, \varphi) &:= \sum_y \mathbb{I}\{\varphi(y) \neq m\} P_{Y|X, \Theta}(y|f(m), \theta), \\ e(\mathcal{V}, \psi) &:= \max_{\theta} \max_m \sum_{f, \varphi} e_{m, \theta}(f, \varphi) \psi(f, \varphi). \end{aligned} \quad (7)$$

1) *Deterministic code for average probability of error:* For a non-symmetrizable AVC, we have the following result from [8]. It provides for the existence of a deterministic code for average probability of error and gives an upper bound on this error.

Theorem 5.1: Let $P_X \in \mathcal{P}(\mathcal{X}^n)$ and $Z(x, \bar{x}, y) \in \{0, 1\}$ be a function such that

$$Z(x, \bar{x}, y) Z(\bar{x}, x, y) = 0 \quad \forall x \in \mathcal{X}^n, \bar{x} \in \mathcal{X}^n, y \in \mathcal{Y}^n \quad (8)$$

and $\mathcal{A} \subset \mathcal{X}^n \times \mathcal{Y}^n$ be a typical set. Let $(X, \bar{X}, Y) \sim P_X \times P_{X \times P_{Y|X,\Theta}}$. Then, there exists a deterministic channel code, $f : \mathcal{W} \rightarrow \mathcal{X}^n, \varphi : \mathcal{Y}^n \rightarrow \mathcal{W}$, such that

$$\begin{aligned} & \max_{\theta} \frac{1}{M} \sum_m e_{m,\theta}(f, \varphi) \\ & \leq \sqrt{M^{-1} 2 \ln(3|\mathcal{T}|^n)} + \min_{P_X} \max_{\theta} \left(\mathbb{P}((X, Y) \notin \mathcal{A} | \theta) \right. \\ & \quad \left. + 2M \log e \mathbb{P}(Z(X, \bar{X}, Y) = 0, (X, Y) \in \mathcal{A} | \theta) \right. \\ & \quad \left. + 2 \log 3|\mathcal{T}|^n \max_{\bar{x}} \mathbb{P}(Z(X, \bar{x}, Y) = 0, (X, Y) \in \mathcal{A} | \theta) \right). \end{aligned} \quad (9)$$

2) *Random code for maximum probability of error:* A random code is a pair of random variables, (F, Φ) , taking values in $\{(f, \varphi) \mid f : \mathcal{W} \rightarrow \mathcal{X}^n, \varphi : \mathcal{Y}^n \rightarrow \mathcal{W}\}$ according to some distribution ψ .

Consider a class of decoders for the channel which are defined as follows. Let $h : \mathcal{X}^n \times \mathcal{Y}^n \rightarrow [0, \infty)$ be a non-negative function. For a given y , the decoder φ maps the output y to an element of \mathcal{W} according to the following decision rule.

$$\varphi(y) = \begin{cases} m & \max_{m' \neq m} h(f(m'), y) < h(f(m), y) \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

Let \mathcal{C}_h be the set of all such codes defined as above. Then, for such codes, we have the following result which provides for the existence of a random code with maximum probability of error and gives an upper bound on the probability of error.

Theorem 5.2: Let $(X, Y) \sim P_X \times P_{Y|X,\Theta}$. Let $h(x, y)$ be a non-negative function such that

$$\mathbb{E}[h(X, y)] = 1 \quad \forall y \in \mathcal{Y}^n. \quad (11)$$

Then, there exists a distribution $\psi \in \mathcal{P}(\mathcal{C}_h)$ such that $\forall \epsilon > 0$,

$$e(\mathcal{V}, \psi) \leq \max_{\theta} \mathbb{P}(h(X, Y) < \epsilon^{-1} M \mid \theta) + \epsilon. \quad (12)$$

Proof : From Lemma 12.9 in [6], we have that there exists a distribution $\psi \in \mathcal{P}(\mathcal{C}_h)$, such that $\forall \theta$ and $\forall m \in \mathcal{W}$,

$$\sum_{f, \varphi} e_{m,\theta}(f, \varphi) \psi(f, \varphi) \leq \mathbb{P}(h(X, Y) < \epsilon^{-1} M \mid \theta) + \epsilon. \quad (13)$$

Since the inequality holds for all θ and for all m , taking maximum over θ and m , we get the required inequality. ■

We now construct a stochastic code with maximum probability of error. For this, we require the following lemma which is a slight restatement of random code reduction lemma from [6].

Lemma 5.1: Let $(F, \Phi) \sim \psi$, and let K be a positive integer such that $K > \frac{\log(M|\mathcal{T}|^n)}{e(\mathcal{V}, \psi) - \log(1+e(\mathcal{V}, \psi))}$. Then there exist deterministic channel codes $(f_i, \varphi_i), i = 1, \dots, K$, such that

$$\frac{1}{K} \sum_{i=1}^K e_{m,\theta}(f_i, \varphi_i) < e(\mathcal{V}, \psi) \quad \forall m, \forall \theta. \quad (14)$$

We now construct a channel code $(Q_{X|W}, \varphi_C)$, where $Q_{X|W}$ is a stochastic encoder and φ_C is a deterministic decoder. The construction is according to Theorem 12.13 in [6]. We consider K deterministic codes, $(f_i, \varphi_i)_{i=1}^K$ defined as $f_i : \mathcal{W} \rightarrow \mathcal{X}^n$ and $\varphi_i : \mathcal{Y}^n \rightarrow \mathcal{W}$, satisfying the equation (14). A randomly chosen code from this ensemble is used to communicate the messages from \mathcal{W} . Further, we consider the deterministic code $(\hat{f}, \hat{\varphi})$ defined as $\hat{f} : \{1, \dots, K\} \rightarrow \mathcal{X}^{d_n}$, $\hat{\varphi} : \mathcal{Y}^{d_n} \rightarrow \{1, \dots, K\}$ satisfying the equation (9), where d_n is a function of n . We use

$(\hat{f}, \hat{\varphi})$ to communicate the index i of the chosen code. Consider a random variable $i \in \{1, \dots, K\}$, distributed uniformly and independent of any other random variable. Given $m \in \mathcal{W}$, the encoder chooses the input string $X \in \mathcal{X}^{d_n+n}$ randomly as $X = (\hat{f}(i), f_i(m))$. The decoder decodes $y = (\hat{y}, \bar{y}) \in \mathcal{Y}^{d_n+n}$, where $\hat{y} \in \mathcal{Y}^{d_n}$ and $\bar{y} \in \mathcal{Y}^n$ according to $\varphi_C : \mathcal{Y}^{d_n+n} \rightarrow \mathcal{W}$, where

$$\varphi_C(y) = \begin{cases} m & \text{if } (\hat{\varphi}(\hat{y}), \varphi_i(\bar{y})) = (i, m) \text{ for some } i \\ 0 & \text{else} \end{cases}$$

For the above defined pair of code $(Q_{X|W}, \varphi_C)$, we have the following upper bound on the probability of error.

Let \mathcal{A} and $Z(x_a, \bar{x}_a, y_a)$ are as defined in Theorem 5.1 and $h(x_a, y_a)$ is a non-negative function satisfying (11). The upper bound on the probability of error is given as follows.

Theorem 5.3: Let $W \sim P_W \in \mathcal{P}(\mathcal{W})$ and $\epsilon > 0$. Then the error $\mathcal{E}_C = \mathcal{E}_C(Q_{X|W}, \varphi_C)$ in code, $(Q_{X|W}, \varphi_C)$, satisfies $\mathcal{E}_C :=$

$$\begin{aligned} & \max_{\theta} \sum_{w, x, y} \mathbb{I}\{\varphi_C(y) \neq w\} P_W(w) Q_{X|W}(x|w) P_{Y|X,\Theta}(y|x, \theta) \\ & \leq \sqrt{K^{-1} 2 \ln(3|\mathcal{T}|^{d_n})} + \min_{P_{X_a}} \max_{\theta_a} \left[\mathbb{P}((X_a, Y_a) \notin \mathcal{A} | \theta_a) \right. \\ & \quad \left. + 2K \log e \mathbb{P}(Z(X_a, \bar{X}_a, Y_a) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \right. \\ & \quad \left. + \max_{\bar{x}_a} 2 \log 3|\mathcal{T}|^{d_n} \mathbb{P}(Z(X_a, \bar{x}_a, Y) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \right] \\ & \quad + \min_{P_{X_b}} \max_{\theta_b} \mathbb{P}(h(X_b, Y_b) < \epsilon^{-1} M \mid \theta_b) + \epsilon, \end{aligned} \quad (15)$$

where $\mathcal{X}^{d_n} \ni \bar{X}_a \sim P_{X_a}$, $\mathcal{X}^{d_n} \times \mathcal{Y}^{d_n} \ni (X_a, Y_a) \sim P_{X_a} P_{Y|X,\Theta=\theta_a}$, with $\theta_a \in \mathcal{T}^{d_n}$, $\mathcal{X}^n \times \mathcal{Y}^n \ni (X_b, Y_b) \sim P_{X_b} P_{Y|X,\Theta=\theta_b}$, with $\theta_b \in \mathcal{T}^n$ and $P_{X_a} \in \mathcal{P}(\mathcal{X}^{d_n})$ and $P_{X_b} \in \mathcal{P}(\mathcal{X}^n)$ are any distributions.

Proof : Let $\widehat{W} = \varphi_C(Y)$. Clearly, \mathcal{E}_C can be written as,

$$\begin{aligned} \mathcal{E}_C &= \max_{\theta} \frac{1}{K} \sum_{i=1}^K \mathbb{P}(\widehat{W} \neq W | i = i, \Theta = \theta) \\ &= \max_{\theta} \left(\frac{1}{K} \sum_{i=1}^K \mathbb{P}(\widehat{W} \neq W, \hat{\varphi}(Y_a) \neq i | i = i, \Theta = \theta) \right. \\ & \quad \left. + \frac{1}{K} \sum_{i=1}^K \mathbb{P}(\widehat{W} \neq W, \hat{\varphi}(Y_a) = i | i = i, \Theta = \theta) \right) \\ &\leq \max_{\theta_a} \frac{1}{K} \sum_{i=1}^K \mathbb{P}(\hat{\varphi}(Y_a) \neq i | i = i, \theta_a) \\ & \quad + \max_{\theta_b} \frac{1}{K} \sum_{i=1}^K \sum_{w=1}^M P_W(w) \mathbb{P}(\varphi_i(Y_b) \neq w | \theta_b) \\ &= \max_{\theta_a} \frac{1}{K} \sum_{i=1}^K e_{i,\theta_a}(\hat{f}, \hat{\varphi}) \\ & \quad + \max_{\theta_b} \frac{1}{K} \sum_{i=1}^K \sum_{w=1}^M P_W(w) e_{w,\theta_b}(f_i, \varphi_i). \end{aligned}$$

Using Theorem 5.1, Lemma 5.1 and Theorem 5.2, we get

$$\mathcal{E}_C \leq \sqrt{K^{-1} 2 \ln(3|\mathcal{T}|^{d_n})} + \max_{\theta_a} \left[\mathbb{P}((X_a, Y_a) \notin \mathcal{A} | \theta_a) \right]$$

$$\begin{aligned}
& + 2K \log e \mathbb{P}(Z(X_a, \bar{X}_a, Y_a) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\
& + \max_{\bar{x}_a} 2 \log 3 |\mathcal{T}|^{d_n} \mathbb{P}(Z(X_a, \bar{x}_a, Y) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\
& + \max_{\theta_b} \mathbb{P}(h(X_b, Y_b) < \epsilon^{-1} M | \theta_b) + \epsilon. \tag{16}
\end{aligned}$$

Taking minimum over the distributions P_{X_a} and P_{X_b} , the result follows. \blacksquare

B. Source code

We now state an upper bound for the probability of error for the source code. The following result is from Lemma 1.3.1 in [9].

Theorem 5.4: Let M be any positive integer. Then, there exists an source code (f_S, φ_S) such that the error $\mathcal{E}_S = \mathcal{E}_S(f_S, \varphi_S)$ satisfies

$$\mathcal{E}_S := \sum_s \mathbb{I}\{\varphi_S \circ f_S(s) \neq s\} P_S(s) \leq \mathbb{P}(-\log P_S(S) \geq \log M).$$

C. Source-channel code and upper bound

We now use deterministic source code and the stochastic channel code from the earlier analysis to derive an upper bound for the minimax value of the game.

Theorem 5.5: The minimax value of the game, $\bar{\nu}(k, d_n + n)$, is bounded above as

$$\begin{aligned}
\bar{\nu}(k, d_n + n) & \leq \min_M \left(\mathbb{P}(-\log P_S(S) \geq \log M) \right. \\
& + \sqrt{K^{-1} 2 \ln(3|\mathcal{T}|^{d_n})} + \min_{P_{X_a}} \max_{\theta_a} \left[\mathbb{P}((X_a, Y_a) \notin \mathcal{A} | \theta_a) \right. \\
& + 2K \log e \mathbb{P}(Z(X_a, \bar{X}_a, Y_a) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\
& + \max_{\bar{x}_a} 2 \log 3 |\mathcal{T}|^{d_n} \mathbb{P}(Z(X_a, \bar{x}_a, Y) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\
& \left. \left. + \min_{P_{X_b}} \max_{\theta_b} \mathbb{P}(h(X_b, Y_b) < \epsilon^{-1} M | \theta_b) \right] + \epsilon, \tag{17}
\right)
\end{aligned}$$

where X_a, X_b, Y_a, Y_b are as in Theorem 5.3 and $\epsilon > 0$.

Proof : The maximum probability of error is bounded as

$$\begin{aligned}
\max_{\theta} \mathbb{P}(\hat{S} \neq S | \Theta = \theta) & \leq \sum_s \mathbb{I}\{\varphi_S \circ f_S(s) \neq s\} P_S(s) + \\
\max_{\theta} \sum_{w,x,y} \mathbb{I}\{\varphi_C(y) \neq w\} P_W(w) Q_{X|W}(x|w) P_{Y|X,\Theta}(y|x, \theta) \\
& = \mathcal{E}_S(f_S, \varphi_S) + \mathcal{E}_C(Q_{X|W}, \varphi_C). \tag{18}
\end{aligned}$$

Using $(Q_{X|W}, \varphi_C)$ and (f_S, φ_S) as in Theorem 5.1 and Theorem 5.4 respectively, we get

$$\begin{aligned}
\max_{\theta} \mathbb{P}(\hat{S} \neq S | \Theta = \theta) & \leq \mathbb{P}(-\log P_S(S) \geq \log M) \\
& + \sqrt{K^{-1} 2 \ln(3|\mathcal{T}|^{d_n})} + \min_{P_{X_a}} \max_{\theta_a} \left[\mathbb{P}((X_a, Y_a) \notin \mathcal{A} | \theta_a) \right. \\
& + 2K \log e \mathbb{P}(Z(X_a, \bar{X}_a, Y_a) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\
& + \max_{\bar{x}_a} 2 \log 3 |\mathcal{T}|^{d_n} \mathbb{P}(Z(X_a, \bar{x}_a, Y) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\
& \left. + \min_{P_{X_b}} \max_{\theta_b} \mathbb{P}(h(X_b, Y_b) < \epsilon^{-1} M | \theta_b) + \epsilon. \tag{19}
\right]
\end{aligned}$$

Thus, taking minimum over M and in the above inequality, we get the required result. \blacksquare

VI. ASYMPTOTIC ANALYSIS

We define the following information quantities.

$$\begin{aligned}
i(x; y) &= \log \frac{(q_\Theta P_{Y|X,\Theta})(y|x)}{(P_X q_\Theta P_{Y|X,\Theta})(y)} \\
V_+ &= \min_{P_X \in \mathcal{P}(\mathcal{X})^*} \max_{q_\Theta \in \mathcal{P}(\mathcal{T})^*} \mathbb{E}[(i(\mathbb{X}; \mathbb{Y}) - \mathbb{E}[i(\mathbb{X}; \mathbb{Y})|\mathbb{X}] \\
&\quad - \mathbb{E}[i(\mathbb{X}; \mathbb{Y})|\Theta] + \mathbb{E}[i(\mathbb{X}; \mathbb{Y})]^2)^2] \tag{20}
\end{aligned}$$

where $(\mathbb{X}, \Theta, \mathbb{Y}) \sim P_{\mathbb{X}} \times q_\Theta \times P_{Y|X,\Theta}$, $(q_\Theta P_{Y|X,\Theta})(y|x) := \sum_{\theta \in \mathcal{T}} q_\Theta(\theta) P_{Y|X,\Theta}(y|x, \theta)$, $(P_X q_\Theta P_{Y|X,\Theta})(y) := \sum_{x \in \mathcal{X}, \theta \in \mathcal{T}} P_{\mathbb{X}}(x) q_\Theta(\theta) P_{Y|X,\Theta}(y|x, \theta)$ and $\mathcal{P}(\mathcal{X})^*$, $\mathcal{P}(\mathcal{T})^*$ are the set of distributions that achieve the optimal in (3).

Take $\bar{P}_{\mathbb{X}_a} \in \mathcal{P}_{d_n}(\mathcal{X})$ such that $\|\bar{P}_{\mathbb{X}_a} - \bar{P}_{\mathbb{X}}\|_\infty \leq \frac{1}{d_n}$, where $\bar{P}_{\mathbb{X}}$ is a capacity achieving distribution that maximizes (20). Define the set $\mathcal{A} \subseteq \mathcal{X}^{d_n} \times \mathcal{Y}^{d_n}$ as

$$\mathcal{A} = \left\{ (x, y) : \log \frac{(q P_{Y|X=x,\Theta})}{(U_{\bar{P}_{\mathbb{X}_a}} q P_{Y|X,\Theta})(y)}(y) > \gamma', q \in \mathcal{P}_{d_n}(\mathcal{S}) \right\},$$

where $(q P_{Y|X=x,\Theta})(y) := \sum_{\theta} q(\theta) P_{Y|X,\Theta}(y|x, \theta)$, $(U_{\bar{P}_{\mathbb{X}_a}} q P_{Y|X,\Theta})(y) := \sum_{x,\theta} U_{\bar{P}_{\mathbb{X}_a}}(x) q(\theta) P_{Y|X,\Theta}(y|x, \theta)$, where $U_{\bar{P}_{\mathbb{X}_a}}$ is the uniform distribution over the type class $T(\bar{P}_{\mathbb{X}_a})$ and $\gamma' = \log(\sqrt{d_n} K |\mathcal{P}_{d_n}(\mathcal{S})|)$. The following lemma which is a trivial modification of the Lemma 6 in [8] uses the non-symmetrizability of the AVC.

Lemma 6.1: Let $P_{\mathbb{X}}(x) > 0$ for all $x \in \mathcal{X}$. Let D_η be the set defined as

$$D_\eta = \{P_{\mathbb{X}\mathbb{X}'\Theta\mathbb{Y}} \in \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{T} \times \mathcal{Y}) : D(P_{\mathbb{X}\mathbb{X}'\Theta\mathbb{Y}} || P_{\mathbb{X}} \times P_{\mathbb{X}'\Theta} \times P_{Y|X,\Theta}) \leq \eta\}, \tag{21}$$

where $(P_{\mathbb{X}} \times P_{\mathbb{X}'\Theta} \times P_{Y|X,\Theta})(x, x', \theta, y) = P_{\mathbb{X}}(x) P_{\mathbb{X}'\Theta}(x', \theta) P_{Y|X,\Theta}(y|x, \theta)$. Let η^* be defined as

$$\eta^* = \inf \{ \eta : P_{\mathbb{X}\mathbb{X}'\Theta\mathbb{Y}} \in D_\eta, P_{\mathbb{X}'\mathbb{X}\Theta'\mathbb{Y}} \in D_\eta \}. \tag{22}$$

If the AVC is non-symmetrizable, then $\eta^* > 0$.

The function $Z : \mathcal{X}^{d_n} \times \mathcal{X}^{d_n} \times \mathcal{Y}^{d_n} \rightarrow \{0, 1\}$ is defined as follows. From Lemma 6.1, we have $\eta^* > 0$. Choose η such that $0 < \eta < \eta^*$. Set $Z(x_a, \bar{x}_a, y_a) = 1$ if ' $(x_a, y_a) \in \mathcal{A}$ ' and 'either $(\bar{x}_a, y_a) \notin \mathcal{A}$ or there exists a $\theta_a \in \mathcal{T}^{d_n}$ such that $P_{x_a, \bar{x}_a, \theta_a, y_a} \in D_\eta$ ', otherwise set $Z(x_a, \bar{x}_a, y_a) = 0$. Thus, we have that

$$Z(x_a, \bar{x}_a, y_a) Z(\bar{x}_a, x_a, y_a) = 0 \quad \forall x_a, \bar{x}_a, y_a. \tag{23}$$

If not, then $\exists (x_a, \bar{x}_a, y_a) \in \mathcal{X}^{d_n} \times \mathcal{X}^{d_n} \times \mathcal{Y}^{d_n}$ such that $Z(x_a, \bar{x}_a, y_a) Z(\bar{x}_a, x_a, y_a) = 1$. This implies $(x_a, y_a) \in \mathcal{A}$ and $(\bar{x}_a, y_a) \in \mathcal{A}$ and hence, there exists joint types $P_{x_a, \bar{x}_a, \theta_a, y_a} \in D_\eta$ and $P_{\bar{x}_a, x_a, \theta'_a, y_a} \in D_\eta$ for some $\theta_a, \theta'_a \in \mathcal{T}^{d_n}$. Thus, by the definition of η^* , we get $\eta^* \leq \eta$. However, this is a contradiction since η is chosen to be strictly lesser than η^* .

Choose $K = c_0 n$ where $c_0 \in \mathbb{N}$ is such that K satisfies the condition in Lemma 5.1. Let $d_n = \lceil \frac{\log K}{C-\delta} \rceil$, where $\delta > 0$. Notice that $d_n = o(n)$. Set

$$h(x_b, y_b) = \log \frac{(q^* P_{Y_b|X_b=x_b, \Theta_b})(y_b)}{(P_{X_b}^* q^* P_{Y_b|X_b, \Theta_b})(y_b)}, \tag{24}$$

where $(q^* P_{Y_b|X_b=x_b, \Theta_b})(y_b) := \sum_{\theta_b} q^*(\theta_b) P_{Y_b|X_b, \Theta_b}(y_b | x_b, \theta_b)$, $(P_{X_b}^* q^* P_{Y_b|X_b, \Theta_b})(y_b) :=$

$\sum_{x_b, \theta_b} P_{X_b}^*(x_b) q^*(\theta_b) P_{Y|X, \Theta}(y_b|x_b, \theta_b)$, $q^*(\theta_b) = \prod_{i=1}^n q_\Theta^*(\theta_i)$, $P_{X_b}^*(x_b) = \prod_{i=1}^n P_{\mathbb{X}}^*(x_i)$, with $q_\Theta^* \in \mathcal{P}(\mathcal{T})$ and $P_{\mathbb{X}}^* \in \mathcal{P}(\mathcal{X})$ being some capacity achieving distributions.

Theorem 6.1: Consider a sequence of (k, n) such that $\lim_{k, n \rightarrow \infty} \frac{k}{d_n + n} < \frac{C}{H(\mathbb{S})}$. Then, $\bar{\nu}(k, d_n + n) \rightarrow 0$ and $\underline{\nu}(k, d_n + n) \rightarrow 0$.

Proof : Fix $P_{X_a}(x_a) = \prod_{i=1}^{d_n} \bar{P}_{\mathbb{X}_a}(x_i)$, where $\bar{P}_{\mathbb{X}_a}$ is defined as earlier, $P_{X_b}(x_b) = \prod_{i=1}^n P_{\mathbb{X}}^*(x_i)$. Consider the sequence (k, n) such that $\frac{k}{d_n + n} < \frac{C - \delta}{H(\mathbb{S}) + \delta}$. Take $M(k, n)$ such that $k(H(\mathbb{S}) + \delta) < \log M(k, n) < (d_n + n)(C - \delta)$.

Using $M(k, n) > 2^{k(H(\mathbb{S}) + \delta)}$ and law of large numbers, we have

$$\mathbb{P}(-\log P_S(S) > k(H(\mathbb{S}) + \delta)) \rightarrow 0. \quad (25)$$

For the chosen Z and \mathcal{A} , we have the following inequality from Theorem 3 in [8].

$$\begin{aligned} & \sqrt{K^{-1} 2 \ln(3|\mathcal{T}|^{d_n})} + \max_{\theta_a} \left[\mathbb{P}((X_a, Y_a) \notin \mathcal{A} | \theta_a) \right. \\ & + 2K \log e \mathbb{P}(Z(X_a, \bar{X}_a, Y_a) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \\ & \left. + \max_{\bar{x}_a} 2 \log 3|\mathcal{T}|^{d_n} \mathbb{P}(Z(X_a, \bar{x}_a, Y) = 0, (X_a, Y_a) \in \mathcal{A} | \theta_a) \right] \\ & \leq Q \left(\sqrt{\frac{d_n}{V_+}} \left(C - \frac{\log K}{d_n} - O\left(\frac{\log d_n}{d_n}\right) \right) \right) + O\left(\frac{1}{\sqrt{d_n}}\right), \end{aligned} \quad (26)$$

where Q is the complementary Gaussian function. It goes to 0 as $d_n \rightarrow \infty$ since $\log K = d_n(C - \delta)$.

We have $\frac{\log M}{n} < (C - \delta)(1 + \frac{d_n}{n}) < C - \delta + (C - \delta)\frac{d_n}{n}$. Since $\frac{d_n}{n} \approx \frac{\log c_0 n}{(C - \delta)n} \rightarrow 0$, we have $\frac{\log M}{n} < C$ for large enough n . Thus, from Lemma 12.10 in [6] we get that, as $n \rightarrow \infty$.

$$\max_{\theta_b} \mathbb{P}(h(X_b, Y_b) < \epsilon^{-1} M | \theta_b) \rightarrow 0. \quad (27)$$

Thus, from (25), (26) and (27), we get $\bar{\nu}(k, d_n + n) \leq \epsilon$ as $k, n \rightarrow \infty$ for $\lim_{k, n} \frac{k}{d_n + n} < \frac{C - \delta}{H(\mathbb{S}) + \delta}$ and $\frac{d_n}{n} \rightarrow 0$. Since δ and ϵ can be arbitrarily small, the result follows. \blacksquare

From (5), it suffices to construct $q, P_{\bar{Y}_q}$ and γ to get a lower bound on $\underline{\nu}(k, n)$. Take $q(\theta) = q^*(\theta) = \prod_{i=1}^n q_\Theta^*(\theta_i)$, where q_Θ^* is a capacity achieving distribution, $P_{\bar{Y}_q}(y) = (P_X^* q^* P_{Y|X, \Theta})(y) := \sum_{x, \theta} P_X^*(x) q^*(\theta) P_{Y|X, \Theta}(y|x, \theta)$, $P_X^*(x) = \prod_{i=1}^n P_{\mathbb{X}}^*(x_i)$, where $P_{\mathbb{X}}^*$ is a capacity achieving distribution and $\gamma = k\alpha$ with $\alpha > 0$,

Theorem 6.2: Consider a sequence of (k, n) such that $\lim_{k, n \rightarrow \infty} \frac{k}{n} > \frac{C}{H(\mathbb{S})}$. Then, $\underline{\nu}(k, n) \rightarrow 1$ and $\bar{\nu}(k, n) \rightarrow 1$.

Proof : Consider a sequence (k, n) , such that $\frac{k}{n} > \frac{C}{H(\mathbb{S}) - 3\alpha}$, where $\alpha > 0$. Let $\bar{i}(x; y) := \log \frac{(q^* P_{Y|X, \Theta})(y|x)}{(P_X^* q^* P_{Y|X, \Theta})(y)}$. Using Theorem 4.1, we get

$$\begin{aligned} \underline{\nu}(k, n) & \geq \max_q \text{OPT}(\text{DP}(q)) \geq \sum_s P_S(s) \left[-\frac{\exp(-k\alpha)}{P_S(s)} \right. \\ & \left. + \min_x P_{Y_q^*|X=x} \{ \bar{i}(x; Y) \leq -\log P_S(s) - k\alpha \} \right] \end{aligned}$$

$$\begin{aligned} & \geq \sum_s P_S(s) \left[\min_x P_{Y_q^*|X=x} \left\{ \bar{i}(x; Y) \leq nC + k\alpha \right\} \right. \\ & \left. \times \mathbb{I}\{-\log P_S(s) \geq nC + 2k\alpha\} - \frac{\exp(-k\alpha)}{P_S(s)} \right] \end{aligned} \quad (28)$$

$$\begin{aligned} & \geq \min_x P_{Y_q^*|X=x} \left\{ \bar{i}(x; Y) \leq \sum_{i=1}^n \mu_i + n\alpha' \right\} \\ & \times \mathbb{P}(-\log P_S(S) \geq kH(\mathbb{S}) - k\alpha) - \exp(-k\alpha), \end{aligned} \quad (29)$$

where $\mu_i := \mathbb{E} \left[\log \frac{(q_\Theta^* P_{Y|X, \Theta})(\mathbb{Y}_i|x_i)}{(P_X^* q_\Theta^* P_{Y|X, \Theta})(\mathbb{Y}_i)} \right]$ with $\mathbb{Y}_i \sim (q_\Theta^* P_{Y|X=x_i, \Theta})$ and $\alpha' := \frac{C\alpha}{H(\mathbb{S}) - 3\alpha}$. (28) follows from $\mathbb{I}\{\bar{i}(x; y) \leq -\log P_S(s) - k\alpha\} \geq \mathbb{I}\{\bar{i}(x; y) \leq nC + k\alpha\} \times \mathbb{I}\{-\log P_S(s) \geq nC + 2k\alpha\}$, (29) follows due to $\mu_i \leq C$ $P_{\mathbb{X}}^*$ -almost surely $\forall i$ and $k(H(\mathbb{S}) - 3\alpha) > nC$. By law of large numbers, the first two terms approach unity as $n \rightarrow \infty$, while $\exp(-k\alpha) \rightarrow 0$ as $k \rightarrow \infty$. Since α can be arbitrarily small, we have for $\lim_{k, n} \frac{k}{n} > \frac{C}{H(\mathbb{S})}$, $\underline{\nu}(k, n) \rightarrow 1$ and hence $\bar{\nu}(k, n) \rightarrow 1$. \blacksquare

VII. CONCLUSION

We formulated the adversarial communication problem as zero-sum game between the encoder-decoder team and the jammer where the encoder-decoder attempt to minimize the probability of error while the jammer tries to maximize it. The problem is non-convex in the space of strategies of the encoder-decoder team and hence a minimax theorem need not hold. However, we showed that an approximate minimax theorem holds for the game. We derived finite blocklength upper and lower bounds for the minimax and maximin values of the game and showed that for rates below $\frac{C}{H(\mathbb{S})}$, the upper and lower values tend to zero and for rates above $\frac{C}{H(\mathbb{S})}$, the values tend to unity. This result is stronger than the strong converse for the joint source-channel coding over AVC and implies the latter.

REFERENCES

- [1] A. A. Kulkarni and T. P. Coleman, "An optimizer's approach to stochastic control problems with nonclassical information structures," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 937–949, 2015.
- [2] S. T. Jose and A. A. Kulkarni, "Linear programming-based converses for finite blocklength lossy joint source-channel coding," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7066–7094, 2017.
- [3] ———, "On a game between a delay-constrained communication system and a finite state jammer," in *Decision and Control (CDC), 2018 IEEE 57th Annual Conference*, to appear.
- [4] I. Csiszar and P. Narayan, "The capacity of the arbitrarily varying channel revisited: Positivity, constraints," *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 181–193, 1988.
- [5] R. Ahlswede, "Elimination of correlation in random codes for arbitrarily varying channels," *Probability Theory and Related Fields*, vol. 44, no. 2, pp. 159–175, 1978.
- [6] I. Csiszar and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.
- [7] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec.*, vol. 4, no. 142-163, p. 1, 1959.
- [8] O. Kosut and J. Kliewer, "Finite blocklength and dispersion bounds for the arbitrarily-varying channel," *arXiv preprint arXiv:1801.03594*, 2018.
- [9] H. Koga *et al.*, *Information-spectrum methods in information theory*. Springer Science & Business Media, 2013, vol. 50.

General Compute and Forward for Virtual Full-Duplex Relaying

Roshan S. Sam

Qualcomm India Private Ltd
Hyderabad 500081, India

Antony V. Mampilly, Srikrishna Bhashyam

Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai 600036, India

Abstract—Motivated by the wireless backhaul application, multihop virtual full duplex relaying using a successive relaying protocol based on compute-and-forward (CoF) was proposed recently by Hong and Caire. The channel gain in each hop was assumed to be equal. In this paper, we consider multihop virtual full duplex relaying where the gain in the different hops can be unequal. We use the recently proposed general compute-and-forward (GCoF) scheme along with successive relaying. GCoF eliminates the non-integer penalty present in CoF or the CoF with simple power allocation used earlier. We determine the achievable rate of virtual full duplex relaying using GCoF for the multihop case and show that this rate is within a constant gap (also independent of the number of hops) of the cutset upper bound under some assumptions.

I. INTRODUCTION

In communication networks, relays are used to improve the network coverage and throughput when the source and destination are far apart and cannot communicate with each other efficiently. The capacity and the corresponding optimal relaying schemes of general relay channels are not known. Among the various relaying schemes, compute-and-forward (CoF) relaying [1] is an important relaying scheme. CoF and the related physical layer network coding strategies were initially studied extensively in the context of two-way relay channels [2], [3].

Recently, motivated by the wireless backhaul application, multihop virtual full duplex relaying using CoF was studied in [4]. In [4], they achieve the performance of a full duplex relay, using two half-duplex relays. They call this relaying scheme as virtual full-duplex relaying. This virtual full-duplex relaying scheme takes advantage of successive relaying as in [5], where the two half-duplex relays alternately transmit in successive slots thereby allowing the source to transmit all the time.

This work was done at the Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, India.

978-1-5386-9286-8/19/\$31.00 ©2019 IEEE

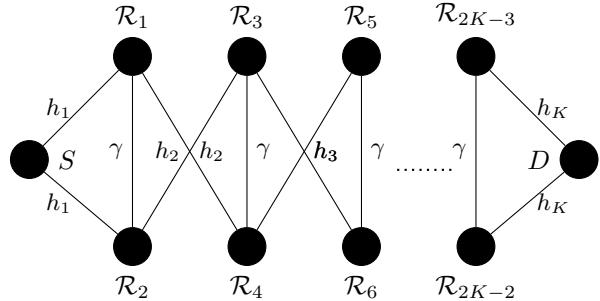


Fig. 1: K -hop network

Consider the K -hop relay network shown in Fig. 1. In [4], they consider the K -hop relay network with $h_1 = h_2 = \dots = h_K = 1$, and each node has the same average transmit power constraint. Some relevant results in [4] are that: (1) successive relaying with dirty paper coding (DPC) achieves the cutset bound for 2-hop relay channel when $\text{SNR} > 1$, (2) CoF with a simple power allocation can achieve rates within a constant gap of the cutset bound for a K -hop relay network with equal hop gains (Fig. 1 with $h_1 = h_2 = \dots = h_K = 1$), and (3) the gap between CoF with power allocation and the cutset bound grows linearly in the number of hops. The cutset bound and gap results in this paper are valid only for the network with equal hop gains. The equal hop gains are assumed to be achieved in the backhaul application by appropriate placement of nodes and power adjustment.

In this paper, we consider the more general K -hop network in Fig. 1. We, therefore, do not require each hop to have the same gain. (Note that this unequal gain network cannot be reduced to a network with equal gains and equal power constraints at all nodes.) We do assume as in [4] that the distance between two relays in the same hop is much smaller than the hop distance, so that the gain for the two relays in each hop is the same. The CoF scheme used in [4] is based on Theorem 1 in [1]. The rates are

found using two methods: with Power Allocation (PA) and without Power Allocation. In this paper, we use the General Compute and Forward (GCoF) formula given in [6] for our protocol. We will denote our scheme as the GCoF scheme. We obtain the following results in this paper: (1) For the 2-hop network, we show that successive relaying with DPC is within 1 bit of the cutset bound under all conditions, (2) We derive an expression for the rate achieved using the GCoF scheme over a K -hop network, (3) We show that the gap between GCoF and the cutset bound is finite as long as the first or last hop is the bottleneck, i.e., $\min\{h_1^2, h_2^2, h_3^2, \dots, h_K^2, \gamma^2\} = h_1^2$ or h_K^2 , and $h_k^2 \text{SNR} > 1, \forall k$. Furthermore, this gap does not grow linearly with the number of hops as in [4].

II. SUCCESSIVE RELAYING WITH DPC

In this section, we consider a 2-hop network. Successive relaying with DPC is known to be not practical for more than 2 hops [4]. We define $h_S \triangleq h_1$ and $h_D \triangleq h_2$. The model for successive relaying is as follows. For odd time slots,

$$\begin{aligned}\underline{\mathbf{y}}_{\mathcal{R}_2}[t] &= h_S \underline{\mathbf{x}}_S[t] + \gamma \underline{\mathbf{x}}_{\mathcal{R}_1}[t] + \underline{\mathbf{z}}_{\mathcal{R}_2}[t], \\ \underline{\mathbf{y}}_D[t] &= h_D \underline{\mathbf{x}}_{\mathcal{R}_1}[t] + \underline{\mathbf{z}}_D[t],\end{aligned}$$

For even time slots,

$$\begin{aligned}\underline{\mathbf{y}}_{\mathcal{R}_1}[t] &= h_S \underline{\mathbf{x}}_S[t] + \gamma \underline{\mathbf{x}}_{\mathcal{R}_2}[t] + \underline{\mathbf{z}}_{\mathcal{R}_1}[t], \\ \underline{\mathbf{y}}_D[t] &= h_D \underline{\mathbf{x}}_{\mathcal{R}_2}[t] + \underline{\mathbf{z}}_D[t],\end{aligned}$$

where $\gamma \in \mathbb{R}$ is the inter-relay interference level and $h_S \in \mathbb{R}$ and $h_D \in \mathbb{R}$ are the channel gains from the source to relay and relay to destination respectively. Here $\underline{\mathbf{x}}_S[t] \in \mathbb{R}^{1 \times n}$ and $\underline{\mathbf{x}}_{\mathcal{R}_k}[t] \in \mathbb{R}^{1 \times n}$ are the signals transmitted by source and relay \mathcal{R}_k . Also, $\underline{\mathbf{y}}_D[t] \in \mathbb{R}^{1 \times n}$ and $\underline{\mathbf{y}}_{\mathcal{R}_k}[t] \in \mathbb{R}^{1 \times n}$ denote the received signals at destination and relay \mathcal{R}_k respectively. Noise is assumed to be i.i.d. Gaussian with zero mean and unit variance (denoted by $\mathcal{N}(0,1)$). Power constraint at each transmitter is denoted by SNR.

The cut set bound for the 2-hop network is given as follows. This is found similar to Section II A in [4].

$$R_{\text{cutset}} = \max_{\substack{t_1, t_2, t_3, t_4 \\ \text{s.t } t_1+t_2+t_3+t_4=1 \\ t_1, t_2, t_3, t_4 \geq 0}} \min\{I_1, I_2, I_3, I_4\} \quad (1)$$

where

$$\begin{aligned}I_1 &\triangleq t_1 C(h_S^2 \text{SNR}) + t_2 C(h_S^2 \text{SNR}) + t_3 C(2h_S^2 \text{SNR}), \\ I_2 &\triangleq t_2 C((h_S^2 + h_D^2 + \gamma^2) \text{SNR} + h_D^2 h_S^2 \text{SNR}^2 + \frac{\gamma^2}{h_D^2}) \\ &\quad + t_3 C(h_S^2 \text{SNR}) + t_4 C(h_D^2 \text{SNR}), \\ I_3 &\triangleq t_1 C((h_S^2 + h_D^2 + \gamma^2) \text{SNR} + h_D^2 h_S^2 \text{SNR}^2 + \frac{\gamma^2}{h_D^2})\end{aligned}$$

$$\begin{aligned}&+ t_3 C(h_S^2 \text{SNR}) + t_4 C(h_D^2 \text{SNR}), \\ I_4 &\triangleq t_1 C(h_D^2 \text{SNR}) + t_2 C(h_D^2 \text{SNR}) + t_4 C(4h_D^2 \text{SNR}),\end{aligned}$$

and $C(x)$ denotes $0.5 \log_2(1+x)$. Here, t_1, t_2, t_3, t_4 is the fraction of time the 2-hop network is in different states. Also, I_1, I_2, I_3, I_4 is the maximum information flow corresponding to the 4 possible cuts of a 2-hop network.

The rate achieved by successive relaying with DPC is given below.

$$R_{\text{DPC}} = \max_{R_{S1}, R_{S2}, R_{1D}, R_{2D}} \frac{1}{2} [\min(R_{S1}, R_{1D}) \\ + \min(R_{S2}, R_{2D})],$$

where R_{Si} denotes the transmission rate from S to \mathcal{R}_i and R_{iD} denotes the transmission rate from \mathcal{R}_i to D .

$$\begin{aligned}\text{If } |h_S| \geq |h_D|, R_{\text{DPC}} &= C(h_D^2 \text{SNR}) \\ \text{If } |h_D| > |h_S|, R_{\text{DPC}} &= C(h_S^2 \text{SNR})\end{aligned}$$

Theorem 1.

$$\begin{aligned}(1) R_{\text{cutset}} - R_{\text{DPC}} &\leq \begin{cases} 1, & \text{if } |h_S| > |h_D| \\ 0.5, & \text{if } |h_S| \leq |h_D| \end{cases} \\ (2) R_{\text{DPC}} &= R_{\text{cutset}} \text{ if and only if } |h_S| = |h_D| = h, \\ &\text{when } |h|^2 \text{SNR} \geq 1\end{aligned}$$

Proof. See the appendices. \square

III. GENERAL COMPUTE AND FORWARD

In this section, we derive the rate achieved using the General Compute and Forward scheme for K -hop virtual full duplex relaying. We denote it by R_{GCoF} . See Fig. 2 for a description of GCoF in each of the time slots. In the first time slot, (See Fig. 2(a)) the source encodes the first message $\underline{\mathbf{w}}_1$ as $\underline{\mathbf{x}}_S(\underline{\mathbf{w}}_1)$ and transmits it to relay 2. The relay 2 decodes $\underline{\mathbf{w}}_1$. In the second time slot, (See Fig. 2(b)) relay 2 re-encodes $\underline{\mathbf{w}}_1$ as $\underline{\mathbf{x}}_{\mathcal{R}}(\underline{\mathbf{u}}_1)$ and transmits it to destination and the destination decodes $\underline{\mathbf{w}}_1$. Also, the source encodes the second message as $\underline{\mathbf{x}}_S(\underline{\mathbf{w}}_2)$ and transmits it to relay 1. The relay 1 receives $\underline{\mathbf{x}}_S(\underline{\mathbf{w}}_2) + \gamma \underline{\mathbf{x}}_{\mathcal{R}}(\underline{\mathbf{u}}_1)$. The relay decodes an integer linear combination of what is sent by the source and the other relay $\underline{\mathbf{u}}_2 = a_2 \underline{\mathbf{w}}_2 + a_1 \underline{\mathbf{u}}_1$. In the third time slot, (See Fig. 2(c)) the relay 1 re-encodes $\underline{\mathbf{u}}_2$ and transmits it as $\underline{\mathbf{x}}_{\mathcal{R}}(\underline{\mathbf{u}}_2)$ to destination. The destination decodes $\underline{\mathbf{w}}_2$ using forward substitution ($\hat{\underline{\mathbf{w}}}_2 = a_2^{-1} \underline{\mathbf{u}}_2 - a_2^{-1} a_1 \underline{\mathbf{u}}_1$). This cycle is continued in successive time slots.

Compared to the CoF scheme in [4], the GCoF coding scheme has additional lattice scaling coefficients that allow us to choose the scaling depending on the channel parameters. This reduces the rate loss

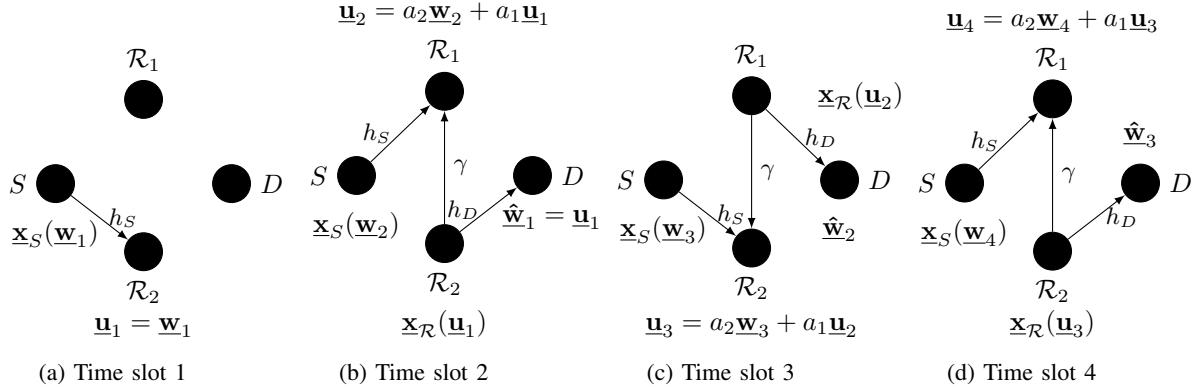


Fig. 2: GCoF for a 2-hop network

due to mismatch between the channel coefficients and the integer combination computed at the relays. Our choice of these scaling coefficients and the integer combination will help us overcome the non-integer penalty.

Theorem 2. *The achievable rate using GCoF protocol for a K-hop network is*

$$R_{GCoF} = \min \left\{ 0.5 \log(1 + h_K^2 \text{SNR}), \right. \\ \left. \min_{i=1,2,\dots,K-1} \left\{ 0.5 \log \left(\frac{h_i^2}{h_i^2 + \gamma^2} + h_i^2 \text{SNR} \right) \right\}, \right. \\ \left. \min_{i=1,2,\dots,K-1} \left\{ 0.5 \log \left(\frac{\gamma^2}{h_i^2 + \gamma^2} + \gamma^2 \text{SNR} \right) \right\} \right\}. \quad (2)$$

Proof. We start by considering the case of a 2-hop network, and later extend the result to a K-hop network.

2-hop network: In Fig. 2(b), the source, relay \mathcal{R}_2 and relay \mathcal{R}_1 form a 2 user Gaussian MAC with channel coefficients h_S and γ respectively. The computation rate tuple of a K-user Gaussian MAC is given in [6, Theorem 1]. Using this result, we note that the relay can reliably decode the linear combination $a_1\mathbf{w}_1 + a_2\mathbf{w}_2$ if

$$R_S \leq 0.5 \log(\beta_1^2) + \\ 0.5 \log \left(\frac{1 + \text{SNR}(h_S^2 + \gamma^2)}{\beta_1^2 a_1^2 + \beta_2^2 a_2^2 + \text{SNR}(h_S \beta_2 a_2 - \gamma \beta_1 a_1)^2} \right), \\ R_{\mathcal{R}} \leq 0.5 \log(\beta_2^2) + \\ 0.5 \log \left(\frac{1 + \text{SNR}(h_S^2 + \gamma^2)}{\beta_1^2 a_1^2 + \beta_2^2 a_2^2 + \text{SNR}(h_S \beta_2 a_2 - \gamma \beta_1 a_1)^2} \right),$$

where, R_S denotes the rate at which the source can transmit and $R_{\mathcal{R}}$ denotes the rate at which the relay can transmit. β_1, β_2 are non zero real numbers which control the rates and $a_1, a_2 \in \mathbb{Z}$. β_1, β_2 and a_1, a_2 can

be chosen to optimize performance. We would like the term $\text{SNR}(h_S \beta_2 a_2 - \gamma \beta_1 a_1)^2$ in the denominator to be zero, so that there is no penalty at high SNR. In order to completely eliminate this loss, we consider $h_S \beta_2 a_2 - \gamma \beta_1 a_1 = 0$. We take $\beta_1^2 a_1^2 + \beta_2^2 a_2^2 = 1$. Solving these two equations, gives:

$$\beta_1 = \frac{h_S}{\sqrt{\gamma^2 + h_S^2}}, \beta_2 = \frac{\gamma}{\sqrt{\gamma^2 + h_S^2}}, a_1 = 1, a_2 = 1.$$

The decoded linear combination at the relay is reliably transmitted to destination if the relay transmits at a rate $R \leq R_2 = 0.5 \log(1 + h_D^2 \text{SNR})$.

Using our choice of β_1, β_2, a_1 , and a_2 above, we get the achievable rate for GCoF as

$$R_{GCoF} = \min\{R_1, R_2\},$$

where $R_1 = \min\{R_S, R_{\mathcal{R}}\}$. After substituting for β_i, a_i we have $R_S = 0.5 \log(1 + \text{SNR}(h_S^2 + \gamma^2)) + 0.5 \log \left(\frac{h_S^2}{h_S^2 + \gamma^2} \right) = 0.5 \log \left(\frac{h_S^2}{h_S^2 + \gamma^2} + h_S^2 \text{SNR} \right)$ and $R_{\mathcal{R}} = 0.5 \log(1 + \text{SNR}(h_S^2 + \gamma^2)) + 0.5 \log \left(\frac{\gamma^2}{h_S^2 + \gamma^2} \right) = 0.5 \log \left(\frac{\gamma^2}{h_S^2 + \gamma^2} + \gamma^2 \text{SNR} \right)$.

K-hop network: Choosing the scaling coefficients at each stage in a similar manner, we get the rate for the general K-hop case to be the expression in (2). The first term in (2) is for the destination to decode the message. Each of the first $K-1$ hops contributes two terms in (2), one for each relay to reliably decode the linear combination. \square

From the above result, we get

$$R_{GCoF} > \min \left[\min_{i=1,2,\dots,K} \{ 0.5 \log(h_i^2 \text{SNR}) \}, \right. \\ \left. 0.5 \log(\gamma^2 \text{SNR}) \right]. \quad (3)$$

For a 2-hop network, when $\min\{h_S^2, h_D^2, \gamma^2\} \neq \gamma^2$, we have

$$R_{GCoF} > \min \{ 0.5 \log(h_S^2 \text{SNR}), 0.5 \log(h_D^2 \text{SNR}) \}.$$

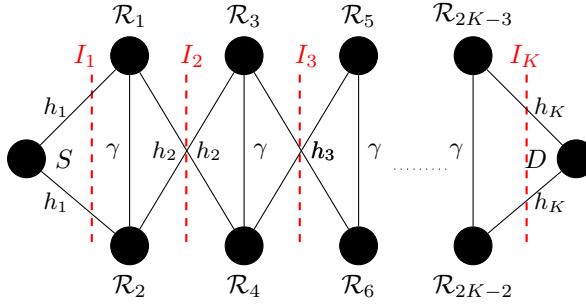


Fig. 3: Cuts for K -hop network

Since we have $R_{DPC} = \min\{0.5 \log(1 + h_S^2 \text{SNR}), 0.5 \log(1 + h_D^2 \text{SNR})\}$, we get $|R_{DPC} - R_{GCoF}| < 0.5$ as long as $h_S^2 \text{SNR}, h_D^2 \text{SNR} \geq 1$.

Corollary 1. For the case when $\text{SNR} \geq 1$, $h_1 = h_2 = \dots = h_K = 1$ and $\gamma > 1$, R_{GCoF} achieves the cutset upper bound within 0.5 bits.

Theorem 3. For a K -hop network, we have the following results.

- If $\min\{h_1^2, h_2^2, h_3^2, \dots, h_K^2, \gamma^2\} = h_1^2$, and $h_1^2 \text{SNR} \geq 1$, then $R_{cutset} - R_{GCoF} \leq 0.5 \log 3$.
- If $\min\{h_1^2, h_2^2, h_3^2, \dots, h_K^2, \gamma^2\} = h_K^2$, and $h_K^2 \text{SNR} \geq 1$, then $R_{cutset} - R_{GCoF} \leq 0.5 \log 5$.

Proof. An upper bound using cutsets, for the K -hop case, can be obtained by considering the cuts as shown in Fig. 3.

$$\begin{aligned} R_{cutset} &= \max \min\{I_1, I_2, I_3, \dots, I_K\} \\ &\leq \min\{\max I_1, \max I_2, \max I_3, \dots, \max I_K\}. \end{aligned}$$

The maximum of I_1 for such a channel is bounded by $C(2h_1^2 \text{SNR})$. Similarly, the maximum of I_k is $2C(h_k^2 \text{SNR})$ for $k = 2, 3, \dots, K-1$. Maximum of I_K is bounded by $C(4h_K^2 \text{SNR})$. Hence, we get

$$R_{cutset} \leq \min\{C(2h_1^2 \text{SNR}), 2C(h_2^2 \text{SNR}), \dots, 2C(h_{K-1}^2 \text{SNR}), C(4h_K^2 \text{SNR})\} \quad (4)$$

Using the lower bound for R_{GCoF} in (3) and the upper bound for R_{cutset} in (4), we bound the gap between the cutset bound and the rate achieved by GCoF for the following cases.

(1) If $\min\{h_1^2, h_2^2, h_3^2, \dots, h_K^2, \gamma^2\} = h_1^2$, then $R_{cutset} \leq C(2h_1^2 \text{SNR})$, $R_{GCoF} > 0.5 \log(h_1^2 \text{SNR})$. Therefore, $R_{cutset} - R_{GCoF} < 0.5 \log\left(\frac{1+2h_1^2 \text{SNR}}{h_1^2 \text{SNR}}\right)$.

If $h_1^2 \text{SNR} \geq 1$, then $R_{cutset} - R_{GCoF} \leq 0.5 \log 3$.

(2) If $\min\{h_1^2, h_2^2, h_3^2, \dots, h_K^2, \gamma^2\} = h_K^2$, then $R_{cutset} \leq C(4h_K^2 \text{SNR})$, $R_{GCoF} > 0.5 \log(h_K^2 \text{SNR})$. Therefore $R_{cutset} - R_{GCoF} < 0.5 \log\left(\frac{1+4h_K^2 \text{SNR}}{h_K^2 \text{SNR}}\right)$.

If $h_K^2 \text{SNR} \geq 1$, then $R_{cutset} - R_{GCoF} \leq 0.5 \log 5$. \square

IV. NUMERICAL EXAMPLES

(1) We consider a 3-hop network where $h_1 \sim \mathcal{N}(2,1)$, $h_2 \sim \mathcal{N}(3,1)$, $h_3 \sim \mathcal{N}(3.5,1)$, and $\gamma \sim \mathcal{N}(5,1)$. Using the realizations where the condition $\min\{h_1^2, h_2^2, h_3^2, \gamma^2\} = h_1^2$ is satisfied, we plot the cutset bound and average achieved rates in Fig. 4. From Fig. 4 we see that GCoF gives significantly better achievable rates than AF (amplify-and-forward) and DF (decode-and-forward) protocols. The AF rate is upper bounded by $R_{AF} < \log\left(1 + \frac{(h_3 h_2 h_1 \beta \beta')^2 \text{SNR}}{\Gamma}\right)$ where $\beta = \sqrt{\frac{\text{SNR}}{1+(h_2^2+\gamma^2)\text{SNR}}}$, $\beta' = \sqrt{\frac{\text{SNR}}{1+(h_1^2+\gamma^2)\text{SNR}}}$ and $\Gamma = 1 + \frac{(h_3 \beta)^2}{1-(\gamma \beta)^2}$. The DF rate is given by $R_{DF} = \min\{0.25 \log(1+(h_1^2+\gamma^2)\text{SNR}), 0.25 \log(1+(h_2^2+\gamma^2)\text{SNR}), 0.5 \log(1+h_3^2 \text{SNR})\}$. In Fig. 5 we plot the histogram of the gap ($= R_{cutset} - R_{GCoF}$). This plot confirms the first part of Theorem 3.

(2) We consider a 3-hop network where $h_1 \sim \mathcal{N}(2,1)$, $h_2 \sim \mathcal{N}(1,1)$, $h_3 \sim \mathcal{N}(4.5,1)$, and $\gamma \sim \mathcal{N}(5,1)$. Using the realizations where the condition $\min\{h_1^2, h_2^2, h_3^2, \gamma^2\} = h_2^2$ is satisfied, we plot the cutset bound and average achieved rates plotted in Fig. 6. From Fig. 6 we see that if the second hop is the limiting hop then the gap between cutset bound and R_{GCoF} can increase with SNR.

(3) We consider a K -hop network with $h_1 = h_2 = \dots = 1$. In [4] the authors consider this network and provide compute-and-forward (CoF), compute-and-forward with power allocation (CoF-PA), AF, DF rates for multihop relay channel. In Fig. 7, we compare these rates with GCoF rate for $\gamma^2 \sim \text{Unif}(2.4, 2.6)$. We observe that GCoF scheme performs significantly better than the CoF, CoF-PA, DF, AF schemes [4, Theorem 2].

V. CONCLUSION

In this paper, we extend the virtual full-duplex relaying scheme in [4] by considering unequal gain across hops for the channel coefficients. We use the General Compute-and-forward (GCoF) scheme to achieve this. We derive an expression for the rate achieved using the GCoF scheme over a K -hop network. We show that the gap between GCoF and the cutset bound is constant as long as the first or last hop is the bottleneck. Furthermore, this gap does not grow linearly with the number of hops as in [4]. We also show that the GCoF

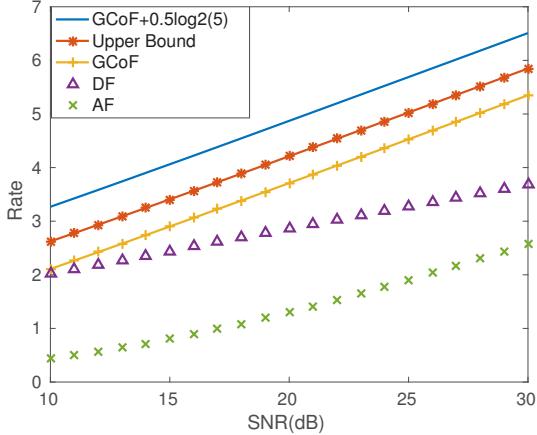


Fig. 4: Average R_{GCoF} vs SNR when $\min\{h_1^2, h_2^2, h_3^2, \gamma^2\} = h_1^2$

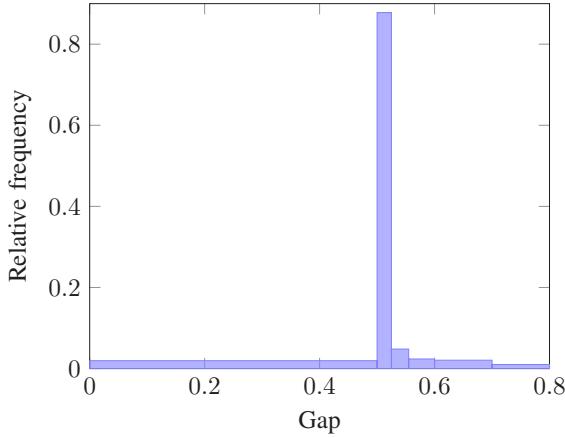


Fig. 5: Histogram of the gap ($= R_{cutset} - R_{GCoF}$) when $\min\{h_1^2, h_2^2, h_3^2, \gamma^2\} = h_1^2$ and SNR = 15dB.

scheme is significantly better than the decode-and-forward and amplify-and-forward schemes. The GCoF scheme is also shown to be significantly better than the compute-and-forward (CoF) and compute-and-forward with power allocation(CoF-PA) proposed in [4] for the unity gain K -hop relay network.

APPENDIX A PROOF OF THEOREM 1(1)

Lemma 1. $\max_{x \in X} \min\{f_1(x), f_2(x)\} \leq \min\{\max_{x \in X} f_1(x), \max_{x \in X} f_2(x)\}$

The cutset bound in (1) can be upper bounded by considering only I_1 and I_4 .

$$R_{cutset} \leq R_{upper} = \max_{t_1+t_2+t_3+t_4=1} \min\{I_1, I_4\}$$

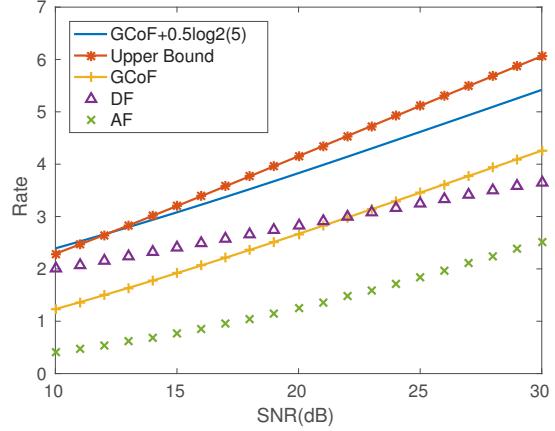


Fig. 6: Average R_{GCoF} vs SNR when $\min\{h_1^2, h_2^2, h_3^2, \gamma^2\} = h_2^2$. Notice that in this case the gap between upper bound and GCoF rate increases with SNR.

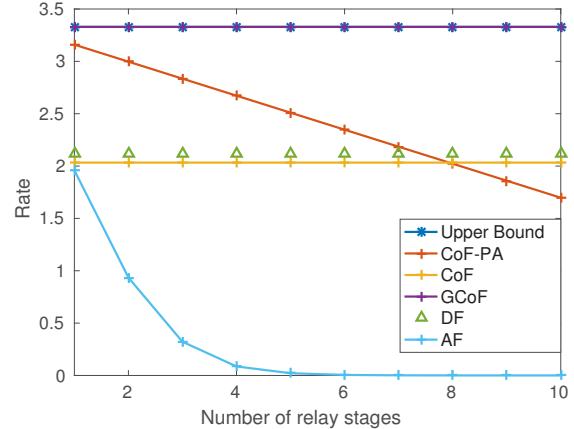


Fig. 7: If $h_1 = h_2 = \dots = 1$ and $\gamma^2 \sim \text{Unif}(2.4, 2.6)$ then the GCoF is significantly better than CoF, CoF-PA, DF, AF schemes in [4]

$$\therefore R_{upper} = \max \min\{C_S \tilde{t} + C_{BC} t_3, C_D \tilde{t} + C_{MAC}(1 - \tilde{t} - t_3)\} \quad (5)$$

where $C_S = C(h_S^2 \text{SNR})$, $C_{BC} = C(2h_S^2 \text{SNR})$, $C_D = C(h_D^2 \text{SNR})$, $C_{MAC} = C(4h_D^2 \text{SNR})$, $\tilde{t} = t_1 + t_2$. In (5), the maximization is over \tilde{t} and t_3 subject to the constraint $\tilde{t} + t_3 \leq 1$. Now by using lemma 1 we have:

$$R_{upper} \leq \min\{x_1, x_2\}$$

where $x_1 = \max C_S \tilde{t} + C_{BC} t_3 = C_{BC}$ ($\because C_S \leq C_{BC}$) and $x_2 = \max C_D \tilde{t} + C_{MAC}(1 - \tilde{t} - t_3)$

$= C_{MAC}$. Hence $R_{upper} \leq \min(C_{BC}, C_{MAC})$.

$$R_{cutset} - R_{DPC} \leq R_{upper} - R_{DPC} \quad (6)$$

If $|h_S| > |h_D|$ then $R_{DPC} = 0.5 \log(1 + h_D^2 \text{SNR})$. Choosing $R_{upper} \leq C_{MAC} = 0.5 \log(1 + 4h_D^2 \text{SNR})$ we can upper bound (6) by 1. If $|h_S| \leq |h_D|$ then $R_{DPC} = 0.5 \log(1 + h_S^2 \text{SNR})$. Choosing $R_{upper} \leq C_{BC} = 0.5 \log(1 + 2h_S^2 \text{SNR})$ we can upper bound (6) by 0.5.

APPENDIX B PROOF OF THEOREM 1(2)

Consider the linear program (LP) to find the cutset bound. It can also be expressed as:

$$\begin{aligned} & \text{Maximize } R_{up} \\ & \text{subject to } R_{up} \leq C_S t_1 + C_{St} t_2 + C_{BCT} t_3 + 0.t_4; \\ & \quad R_{up} \leq 0.t_1 + C_{int} t_2 + C_{St} t_3 + C_{Dt} t_4; \\ & \quad R_{up} \leq C_{int} t_1 + 0.t_2 + C_{St} t_3 + C_{Dt} t_4; \\ & \quad R_{up} \leq C_{Dt} t_1 + C_{Dt} t_2 + 0.t_3 + C_{MAC} t_4; \\ & \quad \sum t_i = 1; t_i \leq 0 \forall i \in [4] \end{aligned} \quad (7)$$

where $C_S = C(h_S^2 \text{SNR})$, $C_{BC} = C(2h_S^2 \text{SNR})$, $C_{int} = C((h_S^2 + h_D^2 + \gamma^2) \text{SNR} + h_D^2 h_S^2 \text{SNR}^2 + \frac{\gamma^2}{h_D^2})$, $C_D = C(h_D^2 \text{SNR})$, $C_{MAC} = C(4h_D^2 \text{SNR})$. The dual program of the above LP is given by:

$$\begin{aligned} & \text{Maximize } R \\ & \text{subject to } R \geq C_S \tau_1 + 0.\tau_2 + C_{int} \tau_3 + C_{Dt} \tau_4; \\ & \quad R \geq C_S \tau_1 + C_{int} \tau_2 + 0.\tau_3 + C_{Dt} \tau_4; \\ & \quad R \geq C_{BCT} \tau_1 + C_S \tau_2 + C_S \tau_3 + 0.\tau_4; \\ & \quad R \geq 0.\tau_1 + C_{Dt} \tau_2 + C_{Dt} \tau_3 + C_{MAC} \tau_4; \\ & \quad \sum_{i=1}^4 \tau_i = 1; \tau_i \leq 0 \forall i \in [4] \end{aligned} \quad (8)$$

In [4] it has been proved that if $|h_S| = |h_D| = 1$ then $t_3^* = t_4^* = 0$. The same proof technique can be easily extended to the case when h_S and h_D are an arbitrary constant i.e. $|h_S| = |h_D| = h$. This proves the necessary part.

Now we need to show the sufficient part i.e. if $t_3^* = t_4^* = 0$ only if $|h_S| = |h_D| = h$. This proof uses the method of contradiction. $t_3^* = t_4^* = 0$ implies $t_1^* = t_2^* = \frac{1}{2}$ because t_1 and t_2 can be swapped in (7) and $\sum t_i = 1$. Substituting $t_1^* = t_2^* = \frac{1}{2}$ and $t_3^* = t_4^* = 0$ in (7) the optimum value of (7) can be obtained.

$$R_{up}^* = \min\{C_S, C_{int}, C_D\} = \min\{C_S, C_D\}$$

The optimal point of (7) is

$$(R_{up}^*, t_1^*, t_2^*, t_3^*, t_4^*) = \left(\min\{C_S, C_D\}, \frac{1}{2}, \frac{1}{2}, 0, 0 \right)$$

At this optimal point we see that the second inequality in (7) is strict. Therefore by complementary slackness we have $\tau_2^* = 0$. Similarly, at the optimal point the third inequality in (7) is strict. Therefore by complementary slackness we have $\tau_3^* = 0$. This implies

$$\tau_1^* + \tau_4^* = 1 \quad (9)$$

Using strong duality theorem we have

$$R^* = R_{up}^* = \min\{C_S, C_D\}.$$

Complementary slackness ensures that the first constraint in (8) should be satisfied by an equality because $t_1^* > 0$ ($t_1^* = \frac{1}{2}$).

$$\begin{aligned} R^* &= C_S \tau_1^* + C_{int} \tau_3^* + C_D \tau_4^* \\ &\Rightarrow \min\{C_S, C_D\} = C_S \tau_1^* + C_D \tau_4^*. \end{aligned} \quad (10)$$

Let $|h_S| > |h_D|$. From (10) we have $C_S \tau_1^* + C_D \tau_4^* = C_D$, which can be simultaneously solved with (9) to get $\tau_1^* = 0$ and $\tau_4^* = 1$. This gives us the optimal point of the dual problem: $(R^*, \tau_1^*, \tau_2^*, \tau_3^*, \tau_4^*) = (C_D, 0, 0, 0, 1)$. But this optimal point doesn't satisfy the fourth inequality of the dual problem (8). So we have a contradiction.

Let $|h_S| < |h_D|$. From (10) we have $C_S \tau_1^* + C_D \tau_4^* = C_S$, which can be simultaneously solved with (9) to get $\tau_1^* = 1$ and $\tau_4^* = 0$. This gives us the optimal point of the dual problem: $(R^*, \tau_1^*, \tau_2^*, \tau_3^*, \tau_4^*) = (C_D, 1, 0, 0, 0)$. But this optimal point doesn't satisfy the third inequality of the dual problem (8). So we have a contradiction.

Since we get contradictions in the cases when $|h_S| > |h_D|$ and $|h_S| < |h_D|$, we conclude that $t_3^* = t_4^* = 0$ only if $|h_S| = |h_D|$.

REFERENCES

- [1] B. Nazer and M. Gastpar, "Compute-and-forward: Harnessing interference through structured codes," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6463–6486, Oct 2011.
- [2] M. P. Wilson, K. Narayanan, H. D. Pfister, and A. Sprintson, "Joint physical layer coding and network coding for bidirectional relaying," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5641–5654, Nov 2010.
- [3] W. Nam, S. Y. Chung, and Y. H. Lee, "Capacity of the gaussian two-way relay channel to within 1/2 bit," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5488–5494, Nov 2010.
- [4] S. N. Hong and G. Caire, "Virtual full-duplex relaying with half-duplex relays," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 4700–4720, Sept 2015.
- [5] W. Chang, S.-Y. Chung, and Y. H. Lee, "Capacity bounds for alternating two-path relay channels," in *Allerton Conference on Communications, Control and Computing, Monticello, IL 2007*, 2007, pp. 1149–1155.
- [6] J. Zhu and M. Gastpar, "Gaussian multiple access via compute-and-forward," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2678–2695, May 2017.

A SegNet Based Image Enhancement Technique for Air-Tissue Boundary Segmentation in Real-Time Magnetic Resonance Imaging Video

Renuka Mannem, Valliappan CA and Prasanta Kumar Ghosh

Electrical Engineering Department, Indian Institute of Science, Bangalore

mannemrenuka@iisc.ac.in, valliappanc@iisc.ac.in, prasantg@iisc.ac.in

Abstract—In this paper, we propose a new technique for segmentation of the Air-Tissue Boundaries (ATBs) in the upper airway of the vocal tract in the midsagittal plane of the real-time Magnetic Resonance Imaging (rtMRI) videos. The proposed technique uses a segmentation using Fisher-discriminant measure (SFDM) scheme. The paper introduces an image enhancement technique using semantic segmentation in the preprocessing of the rtMRI frames before ATB prediction. We use a deep convolutional encoder-decoder architecture (SegNet) for semantic segmentation of the rtMRI images. The paper examines the significance of the preprocessing before ATB prediction by implementing the SFDM approach with different preprocessing techniques. Experiments with 5779 rtMRI video frames from four subjects demonstrate that using the semantic segmentation based image enhancement of rtMRI frames, the performance of the SFDM approach is improved compared to the other preprocessing approaches. Experiment results also show that the proposed approach yields 8.6% less error in ATB prediction compared with a semi-supervised grid based baseline segmentation approach.

Index Terms: air-tissue boundary segmentation, real-time magnetic resonance imaging video, fisher discriminant measure, SegNet, image enhancement.

I. INTRODUCTION

The real-time magnetic resonance imaging video (rtMRI) of the vocal tract in the midsagittal plane during speech is an important tool for speech production research. The rtMRI captures the complete vocal tract in a non-invasive manner [1]. The non-invasive nature of rtMRI makes it more effective than the other existing methods like X-ray [2], Electromagnetic articulography [3] and Ultrasound [4]. The rtMRI video provides the spatio-temporal information of speech articulators which helps in modelling speech production. For this purpose, it is essential to have an accurate Air-Tissue boundary (ATB) segmentation in the rtMRI video. For example, Toutios [5] used the predicted ATBs from the rtMRI video to develop a text-to-speech synthesis system. The rtMRI data is used for comparing the articulatory control of beatboxers to understand the usage of articulators in achieving acoustic goals [6]. The ATB segmentation is used as a pre-processing step in the studies that involve morphological structures of vocal tracts [7] and analysis of vocal tract movement [8] using rtMRI video. The accurate ATB segmentation in the upper airway of the vocal tract is needed to study the time evolution of the vocal tract cross-sectional area [9] which forms the basis for the most speech processing applications. Thus, it is very

important to have an accurate ATB segmentation in the upper airway of the vocal tract in the rtMRI videos before they can be used to study different articulators and dynamics of the vocal tract [10], [11], [12], [13].

The problems of ATB segmentation of rtMRI images have been addressed by several works in the past using various approaches. For example, Asadiabadi et al. presented a statistical method using the appearance and shape model for the vocal tract [14]. Lammert et al. proposed a region of interest (ROI) based technique [15] and a data-driven approach using pixel intensity for the ATB segmentation [16]. A factor analysis approach was used by Toutios et al.[17] and Sorensen et al. [18] to predict the compact outline of the vocal tract. Zhang et al. [19] used multi-directional Sobel operators in order to construct boundary intensity map in the rtMRI video frames. A semantic edge detection based algorithm for contour prediction was proposed by Somandepalli et al. [20]. Several robust ATB segmentation techniques have also been proposed using a composite analysis grid line superimposed on each rtMRI frame [21], [22], [23], [24]. Techniques such as [21], [24], [16], [14] are advantageous over the others because of their unsupervised and semi-automatic approach. However, a more reliable and accurate ATBs can be obtained in a supervised learning approach using the enhanced rtMRI images. For example, Advait et al. proposed a supervised approach using Fisher-discriminant measure (FDM) [26]. Valliappan et al. [25] used a fully convolutional network (FCN) based semantic segmentation with various post-processing steps.

In this paper, we have used FDM based approach [26] for ATB segmentation. The method of ATB segmentation using Fisher-discriminant measure (SFDM) learns the ATBs from the limited training rtMRI frames across different subjects instead of predicting the boundaries using an unsupervised approach. The ATBs, in the upper airway, trace the contours which separate the high pixel intensity regions that correspond to the tissue region from the low pixel intensity regions that correspond to the airway cavity in the vocal tract. Considering the rtMRI images, this transition in the intensity values form air to tissue region is not clearly visible due to the low resolution and blurriness of the images. Hence, the rtMRI images need to be enhanced before applying any ATB segmentation technique to predict reliable boundaries. In this paper, the enhancement of the rtMRI images is achieved by semantically segmenting an image, in which each pixel of the

image is classified into tissue class or airway cavity class. We used the Deep Convolutional Encoder-Decoder Architecture (SegNet) [27] for semantic segmentation of the rtMRI images because of its superior performance compared to the widely used FCN [32]. In the proposed approach, the enhancement of the rtMRI frames using SegNet is done in the preprocessing step before the ATB segmentation using SFDM approach. From the SegNet architecture, two types of output images are extracted: 1) Probability image 2) Binary image. The details of these images are explained in section III-A . The SFDM approach which is implemented using the probability and binary images of SegNet are referred as $FDM_{SegProb}$ and FDM_{SegBin} respectively.

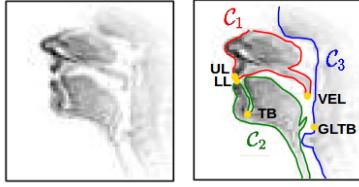


Fig. 1. Illustration of Air-Tissue Boundaries in an rtMRI frame

The performance of the predicted contours is evaluated using the distance between the predicted contours and the manually annotated ground truth contours obtained using Dynamic Time Warping (DTW) [28]. The DTW distances from $FDM_{SegProb}$ and FDM_{SegBin} are compared with a Maeda Grid (MG) based baseline scheme proposed by Kim et al. [21]. The MG scheme uses a grid-based approach to estimate the ATBs within the vocal tract in the midsagittal plane of the rtMRI video frames. The average DTW distances using $FDM_{SegProb}$ and FDM_{SegBin} are found to be 8.5%, 8.6% lesser than that using the baseline MG scheme respectively. The preprocessing of the rtMRI frames before ATB prediction significantly affects the overall performance of the algorithm. To demonstrate the importance of preprocessing, the SFDM algorithm is also implemented with the preprocessing technique used in the baseline MG scheme [21]. The average DTW distances using $FDM_{SegProb}$ and FDM_{SegBin} are found to be 0.87%, 1% lesser than that using the SFDM approach with MG scheme's preprocessing (FDM_{MG}) respectively. And the average DTW distance using $FDM_{SegProb}$ and FDM_{SegBin} are found to be 1.6%, 1.8% lesser than that using the SFDM approach without any preprocessing. The proposed approach can also predict the complete ATBs (both inside and outside vocal tract), unlike the baseline MG approach. The performance of complete ATB segmentation is also reported. The rest of the paper is organized as follows: section II describes the dataset used for experimental analysis. The proposed method of ATB segmentation is explained in section III. Section IV provides a detailed analysis of experimental setup, evaluation metric along with results and section V concludes the paper.

II. DATASET

In this work, we use USC-TIMIT corpus. The USC-TIMIT [29] database consists of the rtMRI videos of the upper

airway in the midsagittal plane, recorded at 23.18 frames/sec. The database contains the videos of five female and five male subjects speaking 460 sentences from MOCHA-TIMIT [30] database. Each frame of the rtMRI video has a spatial resolution of 68×68 pixels ($2.9mm \times 2.9mm$). In this work, we chose to work on 16 rtMRI videos (one for each sentence) from each of two female subjects F1, F2 and two male subjects M1, M2. The selected 16 videos have 1463, 1272, 1642, 1402 frames from subjects F1, F2, M1, M2 respectively. ATBs were drawn manually in each rtMRI frame. A MATLAB based graphical user interface (GUI) was used for manual annotation of the three contours representing the complete ATB in a typical rtMRI frame as shown in Figure 1 . These manually annotated complete ground truth contours are denoted as C_1 , C_2 and C_3 . The details of the manual annotation are available in [31]. Along with the contours, upper lip (UL), lower lip (LL), tongue base (TB), velum tip (VEL) and glottis begin (GLTB) were also marked for each frame using the GUI. As shown in the Figure 1 , the C_1 contour is a closed contour starting from upper lip (UL), through the hard palate and joins the velum (VEL) and goes around the fixed nasal tract. The C_2 contour is also a closed contour which covers the jawline, lower lip (LL), tongue blade and extends below the epiglottis. The C_3 contour marks the pharyngeal wall.

III. PROPOSED METHOD OF ATB PREDICTION

The proposed approach for ATB prediction in the rtMRI images is explained in the Figure 3 . Image enhancement of the rtMRI frames is done in the preprocessing step before contour prediction. For each test rtMRI image, the preprocessing step generates two binary images using the SegNet based semantic segmentation approach for upper and lower contour regions separately. The binary ground truth images corresponding to upper contour region and lower contour region are denoted as $mask_1$ and $mask_2$ respectively. The output binary images from SegNet for upper contour and lower contour regions are denoted as $mask_1^*$ and $mask_2^*$ respectively. Figure 2 illustrates the ground truth and binary images from SegNet architecture for a sample rtMRI image. After preprocessing the test rtMRI frames, the upper and lower contours are predicted separately through SFDM approach using $mask_1^*$ and $mask_2^*$ respectively.

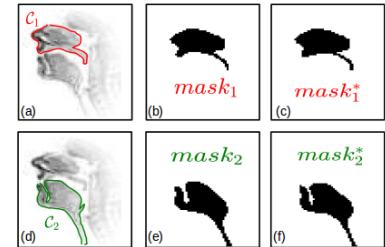


Fig. 2. (a,d) Upper and lower groundtruth contours (b,e) Groundtruth binary images (c,f) Binary images from SegNet based semantic segmentation

A. Preprocessing

Image enhancement of the rtMRI frames across all subjects is done in the preprocessing step before ATB prediction. Se-

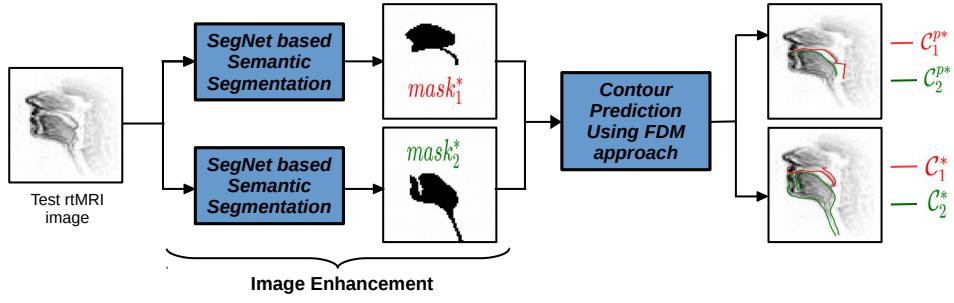


Fig. 3. Illustration of steps in the proposed SegNet based approach

semantic segmentation is utilized to enhance the rtMRI images. The objective of semantic segmentation is pixel classification which associates one of the pre-defined class labels to each pixel of an image. The semantic segmentation generates the images in which the difference between the pixel intensities at the boundary of the tissue and airway cavity region is enhanced. Thus, the semantic segmentation of each rtMRI image helps in estimating precise and reliable ATBs. In this work, we used a deep convolutional encoder-decoder architecture called SegNet [27] for semantic segmentation of the rtMRI images. The better performance of the SegNet is due to its decoder network. The decoder network maps the low resolution feature maps to feature maps of resolution identical to that of the input. It uses the max-pooling indices of the encoder network for non-linear upsampling (Figure 4). The SegNet architecture preserves the spatial information, that is the network takes input image of dimension 68×68 and outputs the image of the same dimension. The encoder network of the SegNet is implemented with the 13 convolutional layered VGG-16 architecture [33]. The SegNet architecture with its encoder and decoder networks is illustrated in the Figure 4.

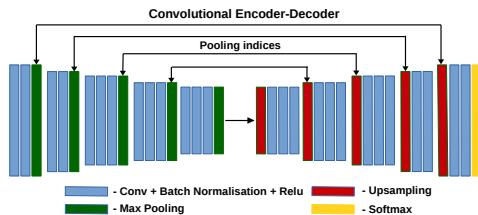


Fig. 4. Block diagram of SegNet architecture with VGG-16

With the limited training data available, the single SegNet model can not learn the two different shapes $mask_1$ and $mask_2$ efficiently. Hence two different SegNet models are employed for learning the upper contour region and lower contour region individually. Each frame of a test rtMRI video is passed through SegNet based semantic segmentation step for each of the two contours (C_1, C_2) separately. For each contour, this step generates a binary image where the class-1 corresponds to the pixels inside the contour region and class-0 corresponds to the pixels outside the contour region. The two different SegNets for each contour (C_1, C_2) are trained as shown in Figure 3. In the testing phase, each frame of

a test rtMRI video is processed using SegNets across all the subjects. The softmax layer in SegNet makes sure that each pixel in the image carries the probability associated with that particular class. An output image of the softmax layer with the pixel intensities as probabilities is referred to as probability image. The binary images are obtained by thresholding these probability images. In this work, we used both binary and probability images from SegNet for contour prediction. Figure 5 illustrates the probability images (with colour map) corresponding to upper and lower contour regions for a sample rtMRI image. In the probability image, the color variation of the pixels from blue to yellow indicates the pixel intensity variation from 0 to 1.

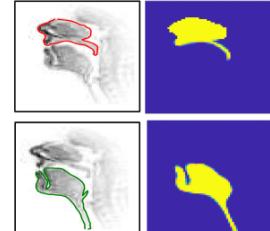


Fig. 5. Probability images for upper (top) and lower (bottom) contours

B. Contour Prediction using SFDM approach

The SFDM approach is used for predicting the contours (C_1, C_2) using the output binary images of SegNet (or the probability images). The ATB prediction in the rtMRI images can be viewed as a problem of finding boundary corresponding to the contour of maximal contrast. Thus, in the SFDM approach, the ATBs in all frames of the rtMRI video are jointly estimated by maximizing a contrast measure around the predicted ATBs. The SFDM approach uses the FDM as a measure of contrast along the contour for ATB prediction [26]. A temporal smoothness criterion is incorporated to exploit the slowly varying nature of the vocal tract morphology. The criterion predicts the boundaries jointly across multiple video frames which ensures that the ATBs do not change drastically in the consecutive frames. Dynamic programming is used to consolidate the temporal smoothness criterion with the FDM. Using the SFDM approach for ATB segmentation has the advantage of predicting the ATBs outside the vocal tract also which provides a more detailed description of the boundaries in the midsagittal plane of the rtMRI video.

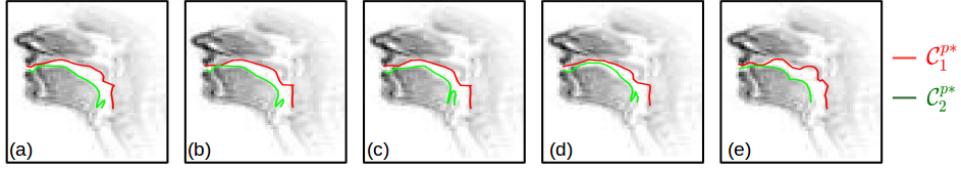


Fig. 6. : Illustration of predicted upper and lower contours (within the vocal tract) from (a) FDM_{SegProb} (b) FDM_{SegBin} (c) FDM_{MG} (d) SFDM (e) MG approaches.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

In this work, The ATBs are predicted from the 16 videos of rtMRI data for each subject (F1, F2, M1, and M2) separately using a four-fold cross-validation setup. In each fold, eight training, four development, and four test videos are used in a round-robin fashion. In the preprocessing step, the SegNet model is trained using the 32 videos (8 training videos from each subject) and tested with 16 videos (4 test videos from each subject). Remaining 16 videos are used as the development set (4 development videos from each subject). Each fold on average consists of ~ 2900 training images, ~ 1443 images in both development and test sets (from all the 4 subjects). The SegNet model is trained for a maximum of 120 epochs with early stopping condition imposed based on the validation loss. As explained in the section III-A, two different SegNet models are trained for upper and lower contours separately. The training contours that correspond to different segments of lower and upper contours are obtained from the manually annotated contours as illustrated in the Figure 1.

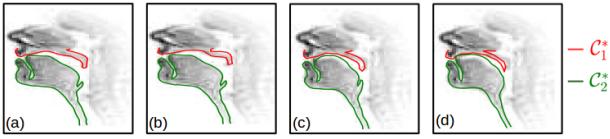


Fig. 7. Illustration of predicted complete upper and lower contours from (a) FDM_{SegProb} (b) FDM_{SegBin} (c) FDM_{MG} (d) SFDM approaches

B. Evaluation metric

For evaluation of the proposed approach, we use two metrics: 1) DTW Distance to measure the alignment between the predicted contour and the ground truth contour [28]. The DTW scores have a unit of pixel. 2) Pixel accuracy to evaluate the SegNet architecture's performance in semantic segmentation [27].

1) Dynamic Time Warping distance: The DTW distance provides a measure of alignment between two contours through an optimal map between the points of the given contours [28]. Lets denote the ground truth contour C_g of length L_g as $C_g(i) = \{(x_i^g, y_i^g) | 1 \leq i \leq L_g\}$ and predicted contour C_p of length L_p as $C_p(i) = \{(x_i^p, y_i^p) | 1 \leq i \leq L_p\}$. The DTW distance between C_g and C_p is defined as:

$$\mathcal{M}_{dtw}(C_g, C_p) = \min_{\substack{1 \leq m'_g \leq L_g \\ 1 \leq m'_p \leq L_p}} \frac{1}{L} \sum_{l=1}^L \|C_g(m'_g(l)) - C_p(m'_p(l))\|_2 \quad (1)$$

If the contours C_p and C_g have similar shape and positioned close to one another, then the DTW distance $\mathcal{M}_{dtw}(C_g, C_p)$ is small. From the above equations, it can be seen that the value of L depends on the lengths of the contours C_p and C_g (L_p and L_q respectively). In this work, two types of performance evaluations are done: (1) evaluation of the complete predicted ATBs C_1^*, C_2^* . The results of the proposed approaches FDM_{SegProb} and FDM_{SegBin} are compared with the SFDM approach using MG scheme's preprocessing (FDM_{MG}) and without using any preprocessing (SFDM). (2) evaluation of the predicted ATBs within the vocal tract C_1^{p*}, C_2^{p*} . The results of the proposed methods are compared with the baseline MG, FDM_{MG} and SFDM approaches. In order to obtain the evaluations for C_1^{p*}, C_2^{p*} contours, the complete ground truth upper and lower contours (C_1, C_2) are pruned using the contour pruning method followed in [26]. The pruned upper and lower ground truth contours are denoted as C_1^p, C_2^p respectively.

2) Pixel Accuracy: In order to evaluate the performance of the SegNet architecture used for the semantic segmentation, pixel accuracy is used. Pixel accuracy indicates the fraction of the pixels that are correctly classified in the output binary image of SegNet compared to the ground truth image. Let p_{ij} be the number of pixel of class i predicted to class j and L_i is total number of pixels in class i , i.e., $L_i = \sum_j p_{ij}$, where $i, j \in \{0, 1\}$. Then the pixel accuracy is defined as $\frac{\sum_i p_{ii}}{\sum_i L_i}$.

Lower Contour				
Sub	FDM _{SegProb}	FDM _{SegBin}	FDM _{MG}	SFDM
F1	0.804 ± 0.117	0.812 ± 0.121	0.834 ± 0.121	0.799 ± 0.119
F2	0.962 ± 0.172	0.945 ± 0.164	0.957 ± 0.168	0.998 ± 0.191
M1	0.940 ± 0.209	0.940 ± 0.204	0.961 ± 0.150	0.966 ± 0.168
M2	0.959 ± 0.209	0.963 ± 0.217	0.998 ± 0.332	1.015 ± 0.305
Avg	0.915 ± 0.178	0.914 ± 0.177	0.937 ± 0.191	0.943 ± 0.194

TABLE I
DTW DISTANCE (AVERAGE \pm STANDARD DEVIATION) OF THE COMPLETE LOWER CONTOURS USING FDM_{SegProb}, FDM_{SegBin}, FDM_{MG} AND SFDM

Upper Contour				
Sub	FDM _{SegProb}	FDM _{SegBin}	FDM _{MG}	SFDM
F1	0.873 ± 0.096	0.872 ± 0.099	0.90 ± 0.121	0.905 ± 0.119
F2	1.038 ± 0.165	1.044 ± 0.167	1.061 ± 0.170	1.060 ± 0.172
M1	1.073 ± 0.177	1.076 ± 0.181	1.070 ± 0.180	1.068 ± 0.184
M2	1.090 ± 0.208	1.087 ± 0.206	1.102 ± 0.240	1.100 ± 0.238
Avg	1.019 ± 0.162	1.020 ± 0.164	1.033 ± 0.177	1.033 ± 0.178

TABLE II
DTW DISTANCE (AVERAGE \pm STANDARD DEVIATION) OF THE COMPLETE UPPER CONTOURS USING FDM_{SegProb}, FDM_{SegBin}, FDM_{MG} AND SFDM

C. Results and Discussions

Table I and II show the average (\pm standard deviation) of $\mathcal{M}_{dtw}(C_1, C_1^*)$ and $\mathcal{M}_{dtw}(C_2, C_2^*)$ (for complete ATBs)

Upper Contour					
Sub	FDM _{SegProb}	FDM _{SegBin}	FDM _{MG}	SFDM	MG
F1	0.917±0.119	0.917±0.121	0.940±0.167	0.960±0.159	1.023±0.191
F2	1.147±0.196	1.159±0.199	1.157±0.194	1.160±0.120	1.246±0.292
M1	1.118±0.206	0.120±0.207	1.112±0.201	1.113±0.207	1.110±0.208
M2	1.117±0.237	1.113±0.234	1.106±0.230	1.108±0.234	1.192±0.245
Avg	1.073±0.189	1.075±0.190	1.077±0.198	1.084±0.200	1.132±0.227

TABLE III

DTW DISTANCE (AVERAGE ± STANDARD DEVIATION) OF THE UPPER CONTOURS (WITHIN THE VOCAL TRACT) USING FDM_{SegProb}, FDM_{SegBin}, MG, FDM_{MG} AND SFDM

Lower Contour					
Sub	FDM _{SegProb}	FDM _{SegBin}	FDM _{MG}	SFDM	MG
F1	0.963±0.198	0.968±0.206	0.998±0.234	0.946±0.208	1.214±0.213
F2	1.218±0.249	1.189±0.243	1.240±0.248	1.261±0.288	1.283±0.275
M1	1.168±0.244	1.172±0.249	1.170±0.257	1.211±0.278	1.263±0.640
M2	1.156±0.275	1.149±0.268	1.163±0.409	1.190±0.386	1.358±0.305
Avg	1.124±0.241	1.119±0.246	1.141±0.287	1.150±0.289	1.274±0.356

TABLE IV

DTW DISTANCE (AVERAGE ± STANDARD DEVIATION) OF THE LOWER CONTOURS (WITHIN THE VOCAL TRACT) USING FDM_{SegProb}, FDM_{SegBin}, MG, FDM_{MG} AND SFDM

using the FDM_{SegProb}, FDM_{SegBin}, FDM_{MG}, and SFDM approaches and it is observed that the FDM_{SegProb}, FDM_{SegBin} approaches result in lower average DTW distance compared to the FDM_{MG}, SFDM approaches. The average DTW distance using the FDM_{SegProb} is found to be 1.9%, 2.2% lesser than that using the FDM_{MG}, SFDM respectively. And the average DTW distance using the FDM_{SegBin} is found to be 1.8%, 2.1% lesser than that using the FDM_{MG}, SFDM respectively. Figure 7 shows the sample rtMRI frame for which the FDM_{SegProb}, FDM_{SegBin} approaches predicted more accurate complete ATBs than the FDM_{MG}, and SFDM approaches.

Table III and Table IV show the average (\pm standard deviation) of $\mathcal{M}_{dtw}(\mathcal{C}_1^p, \mathcal{C}_1^{p*})$ and $\mathcal{M}_{dtw}(\mathcal{C}_2^p, \mathcal{C}_2^{p*})$ using the FDM_{SegProb}, FDM_{SegBin}, MG, FDM_{MG} and SFDM approaches. From the Table III and Table IV , it is observed that the FDM_{SegProb} and FDM_{SegBin} approaches result in lower average DTW distances compared to the MG, FDM_{MG}, SFDM approaches. The average DTW distance of the lower and upper ATBs in the upper airway across the four subjects from the FDM_{SegProb} approach is 8.5%, 0.87%, 1.6% lesser than that using the MG, FDM_{MG}, and SFDM approaches respectively. The average DTW distance of the lower and upper ATBs in the upper airway across the four subjects from the FDM_{SegBin} approach is 8.6%, 1.01%, 1.8% lesser than that using the MG, FDM_{MG}, and SFDM approaches respectively. Figure 6 shows a sample rtMRI frame for which the FDM_{SegProb} and FDM_{SegBin} approaches predicted the more accurate and reliable ATBs in the upper airway than the MG, FDM_{MG}, and SFDM approaches.

Table V shows the average pixel accuracy (in percentage) using SegNet and FCN models for $mask_1^*$ and $mask_2^*$. From Table V, it is observed that the SegNet model results in a high pixel accuracy for both $mask_1^*$ and $mask_2^*$ which explains the better performance of the SegNet model compared to FCN. On an average $\sim 1\%$ pixels are being misclassified in the output

binary images from SegNet model. These misclassified pixels predominantly lie in the boundary region mainly where the upper and lower ATBs come in contact.

The superior performance of the FDM_{SegProb}, FDM_{SegBin} approaches could be due to the following reasons: 1) The image enhancement of the rtMRI frames in the preprocessing step improves the pixel intensity variation from tissue region to airway cavity region which helps in predicting the precise boundaries, 2) The approach uses the global contrast features of the boundaries using FDM which results in more accurate boundaries, 3) Due to its supervised nature, FDM_{SegProb} and FDM_{SegBin} predict the accurate boundaries by overcoming the imaging artifacts and grainy noise which poses challenges for the unsupervised algorithms.

From the results, it can also be observed that in very few cases, the proposed FDM_{SegProb}, FDM_{SegBin} approaches do not perform better than the other approaches or the performance improvement is insignificant. The poor performance of the FDM_{SegProb}, FDM_{SegBin} approaches could be due to the training of SegNet architecture on the low resolution input rtMRI images. The trained model sometime cannot differentiate the airway cavity and tissue regions perfectly and results in misclassification of pixels. Due to the misclassification, the output binary image of the SegNet model may have undesired high pixel intensity (class-1) regions or low pixel intensity (class-0) regions which eventually affect the precise ATB segmentation.

Sub	$mask_1^*$		$mask_2^*$	
	SegNet	FCN	SegNet	FCN
F1	99.64	99.39	98.59	98.34
F2	99.45	99.20	98.39	98.14
M1	99.53	99.28	98.22	97.97
M2	99.57	99.32	98.34	98.09

TABLE V
AVERAGE PIXEL ACCURACY (IN %) FOR LOWER AND UPPER CONTOUR MASKS FROM SEGNET AND FCN MODELS

V. CONCLUSION

In this paper, we proposed a supervised algorithm using SegNet based image enhancement for the ATB prediction in the midsagittal plane of the rtMRI videos. In addition to this, the paper also highlights the significance of preprocessing of the rtMRI images before contour prediction. The robust performance of the proposed approach is shown to be due to the image enhancement of the rtMRI frames before ATB prediction, considering the global contrast features and the supervised nature of the model. To improve the performance of the proposed approach, we need to investigate on the minimum training data size needed for the SegNet model to achieve high pixel accuracy. Our future work includes predicting the ATBs directly from the neural network instead of using the network for semantic segmentation only.

REFERENCES

- [1] E. Bresch, Y. C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, May 2008.
- [2] D. C. Wold, "Generation of vocal tract shapes from formant frequencies," in *The Journal of the Acoustical Society of America*, vol. 78, no. S1, pp. S54–S55, 1985.
- [3] D. Maurer, B. Grne, T. Landis, G. Hoch, and P. W. Schnle, "Re-examination of the relation between the vocal tract and the vowel sound with electromagnetic articulography in vocalizations," in *Clinical Linguistics & Phonetics*, vol. 7, no. 2, pp. 129–143, 1993.
- [4] K. L. Watkin and J. M. Rubin, "Pseudothree-dimensional reconstruction of ultrasonic images of the tongue," in *The Journal of the Acoustical Society of America*, vol. 85, no. 1, pp. 496–499, 1989.
- [5] A. Toutios, T. Sorensen, K. Somandepalli, R. Alexander, and S. S. Narayanan, "Articulatory synthesis based on real-time magnetic resonance imaging data," in *Interspeech 2016*, pp. 1492–1496.
- [6] N. Patil, T. Greer, R. Blaylock, and S. S. Narayanan, "Comparison of basic beatboxing articulations between expert and novice artists using real-time magnetic resonance imaging," in *Proc. Interspeech 2017*, pp. 2277–2281.
- [7] A. Lammert, M. Proctor, and S. Narayanan, "Interspeaker variability in hard palate morphology and vowel production," in *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 6, pp. 1924–1933, 2013.
- [8] V. Ramanarayanan, L. Goldstein, D. Byrd, and S. S. Narayanan, "An investigation of articulatory setting using real-time magnetic resonance imaging," in *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 510–519, 2013.
- [9] B. H. Story, I. R. Titze, and E. A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.
- [10] B. Parrell and S. Narayanan, "Interaction between general prosodic factors and languagespecific articulatory patterns underlies divergent outcomes of coronal stop reduction," in *International Seminar on Speech Production (ISSP) Cologne, Germany*, 2014.
- [11] F.-Y. Hsieh, L. Goldstein, D. Byrd, and S. Narayanan, "Pharyngeal constriction in English diphthong production," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 060271, 2013.
- [12] A. Prasad, V. Periyasamy, and P. K. Ghosh, "Estimation of the invariant and variant characteristics in speech articulation and its application to speaker identification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4265–4269.
- [13] M. Li, J. Kim, A. Lammert, P. K. Ghosh, V. Ramanarayanan, and S. Narayanan, "Speaker verification based on the fusion of speech acoustics and inverted articulatory signals," in *Computer Speech and Language*, vol. 36, pp. 196 – 211, 2016.
- [14] S. Asadiabadi and E. Erzin, "Vocal tract airway tissue boundary tracking for rtMRI using shape and appearance priors," in *Interspeech*, pp. 636–640, 2017.
- [15] A. C. Lammert, V. Ramanarayanan, M. I. Proctor, S. Narayanan *et al.*, "Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis," in *INTERSPEECH*, 2013, pp. 959–962.
- [16] A. C. Lammert, M. I. Proctor, and S. S. Narayanan, "Data-driven analysis of realtime vocal tract MRI using correlated image regions," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1572–1575.
- [17] A. Toutios and S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data," in *18th International Congress of Phonetic Sciences, ICPHS 2015, Glasgow, UK, August 10-14, 2015*, 2015.
- [18] T. Sorensen, A. Toutios, L. Goldstein, and S. S. Narayanan, "Characterizing vocal tract dynamics across speakers using real-time MRI," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 465–469.
- [19] D. Zhang, M. Yang, J. Tao, Y. Wang, B. Liu, and D. Bukhari, "Extraction of tongue contour in real-time magnetic resonance imaging sequences," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 937–941.
- [20] K. Somandepalli, A. Toutios, and S. S. Narayanan, "Semantic edge detection for tracking vocal tract air-tissue boundaries in real-time magnetic resonance images," in *Interspeech 2017*, pp. 631–635, 2017.
- [21] J. Kim, N. Kumar, S. Lee, and S. S. Narayanan, "Enhanced airway-tissue boundary segmentation for real-time magnetic resonance imaging data," in *International Seminar on Speech Production (ISSP)*, 2014, pp. 222 – 225.
- [22] S. E. Öhman, "Numerical model of coarticulation," in *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967.
- [23] S. Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modelling*. Springer, 1990, pp. 131–149.
- [24] M. I. Proctor, D. Bone, A. Katsamanis, and S. S. Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, 2010, pp. 1576–1579.
- [25] V. CA, R. Mannem, and P. K. Ghosh, "Air-tissue boundary segmentation in real-time magnetic resonance imaging video using semantic segmentation with fully convolutional networks," in *Proc. Interspeech 2018*, 2018, pp. 3132–3136.
- [26] A. Koparkar and P. K. Ghosh, "A supervised air-tissue boundary segmentation technique in real-time magnetic resonance imaging video using a novel measure of contrast and dynamic programming," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5004–5008.
- [27] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," in *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
- [28] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *in The 3rd International Conference on Knowledge Discovery and Data Mining*, ser. AAAIWS'94, 1994, pp. 359–370.
- [29] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein *et al.*, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.
- [30] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar of Speech Production*, 2000, pp. 305–308.
- [31] A. K. Pattem, A. Illa, A. Afshan, and P. K. Ghosh, "Optimal sensor placement in electromagnetic articulography recording for speech production study," in *Computer Speech & Language*, vol. 47, pp. 157–174, 2018.
- [32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.
- [33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

Edge Preserved Herringbone Artifact Removal from MRI Using Two-Stage Variational Mode Decomposition

Divya Pankaj, D. Govind and K. A. Narayanankutty

Center for Computational Engineering and Networking (CEN)

Amrita School of Engineering, Coimbatore

Amrita Vishwa Vidyapeetham, India

Email:d_govind@cb.amrita.edu, p_divyaj@cb.amrita.edu,ka_narayanankutty@yahoo.com

Abstract—Magnetic Resonance Imaging (MRI) is an efficient and non-invasive method for analyzing the structural features and functional behaviors of internal organs and tissues for medical diagnosis. The artifacts present in MRI mislead the diagnostic procedure. Herringbone artifact is a hardware artifact generated from the outlier in k -space measurement. In real-time MRI, the herringbone artifact has non-stationary noise characteristics. The non-stationary noise characteristics affect the high-frequency characteristics which in turn results in an improper estimation of structural details of the image in the processing stage. The objective of the present work is to exploit the properties of the variational mode decomposition (VMD) in reducing the effective herringbone noise at selected spectral regions (high-frequency regions in particular) of the given MRI data. In the present work, the given herringbone artifact affected image is subjected to VMD in two stages. The reconstructed image by removing the higher frequency VMD modes in two-stages found to enhance the noisy MRI data. In the second stage of processing, the discarded higher frequency VMD mode in the first stage is further decomposed into component modes in order to preserve the high-frequency details. The enhanced image is later reconstructed by adding low-frequency modes obtained in both decomposition stages. The effectiveness of the proposed two-stage VMD based enhancement is confirmed from the improved scores obtained from the non-reference quality measures such as Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) and Naturalness Image Quality Evaluator (NIQE).

I. INTRODUCTION

MRI is highly accurate medical imaging technique for medical diagnosis. The anomalies present in the MRI called artifact limits the quality of images which results in a wrong diagnostic procedure. In MRI acquisition, the MR raw data is captured by sampling the echo signal corresponding to k -space. The MRI is reconstructed from the digitized k -space data and each pixel in the image is the resultant of the weighted sum of all the data point in k -space. Thus the information contains in each pixel is associated with entire data points. Therefore, the fluctuations occurred in k -space affected the entire image [1], [2]. The k -space data gets corrupted mainly due to MRI hardware and room shielding, physiological movement of the patient, tissue heterogeneity, presence of foreign bodies, which results in artifacts in MRI [3], [4]. Due to the relevance of MRI in

clinical use, MRI denoising is an open problem for researchers. The noise reduction can preserve the fine structures in MRI which helps the Physicians for an accurate diagnosis. The artifacts observed in MRI are herringbone artifact, magnetic susceptibility artifact, aliasing artifact, motion artifact, eddy current artifact, etc [3]–[6]. Some of these artifacts can be eliminated by the non-mathematical approach with the help of technicians/clinicians. Multiple methodologies are introduced to eliminate MRI artifacts are discussed in [5]–[7] and [4]. Motion artifact occurs due to patient movement or respiratory motion. In [8], Santosh et al. proposed a method based on dynamic mode decomposition for removing motion artifact. Benedikt et al. proposed a movement detection method in MRI using a supervised learning approach based on random decision forests [9]. In [10], Wei et al. used point spread function for correcting bulk in-plane motion artifact in MRI. Rician is a thermal noise usually observed in MRI. A detailed survey on rician denoising is given in [11]. The first approach of MRI denoising was initiated by Henkleman [12], estimated the magnitude from a noisy MRI for a perfect denoising. Rician denoising based on filtering methods are discussed in [13]–[17] and [18].

Herringbone artifact is a hardware artifact occurring in MRI due to the interference of external radio frequency (RF) into the MRI system. This interference generates single or multiple bad points or a spike of noise in k -space. The bad point has low or high intensity compared with the intensity of remaining k -space data. During the reconstruction of the image, this points results in dark stripes in the image. The distance between the center of k -space and the bad point determine the spacing between the stripes. The herringbone noise is also known as criss-cross noise or corduroy artifact or spike noise [1], [4]. The noise is scattered all over the image either in a single plane or multiple planes. The main source of external RF are electromagnetic spikes by the gradient coils, penetration of external noise to scanning room, external electronic devices and electromagnetic materials, static electricity, fluctuation of AC current. Eliminating the source of extrinsic RF and static electricity, avoiding low humidity environment, usage of magnetic resonance compatible equipment and perfectly

closed scanning room are some non-mathematical methods for removing herringbone artifact [4]. Herringbone artifact present in MRI has high-frequency components with non-stationary noise characteristics. Identifying and separating the non-stationary high-frequency component from the image is cumbersome. In order to avoid the repeated MRI scan due to the spike/herringbone noise, several k -space post-processing techniques are introduced. Foo et al. introduced a method for spike noise detection and removal based on suitable data replacement scheme depends on the threshold value [19]. The magnitude of the data point greater than the threshold value is replaced using a data replacement scheme. Kao et al. point out that the thresholding methods are failed to restore the corrupted k -space data, and proposed a Fourier transform method using a window filter [20]. Zhang et al. introduced a noise detection and removal method based on the time course of k -space data. The time course of k -space data is stable, hence the spike detection is performed by observing the time course of each data. The sliding window method is used for removing the spike noise [21]. Huang et al. proposed a partially parallel imaging technique based on k -space convolution for the efficient detection and correction of the herringbone noise. Campbell-Washburn et al. proposed the robust principal component analysis (RPCA) algorithm to remove corrupted k -space data. RPCA decomposed the k -space data into low-rank and sparse components, where the sparse component captured the artifact k -space data and low-rank component captured artifact-free k -space data [22]. Different image space approaches are introduced for removing herringbone artifact. The algorithm based on fast Fourier transform (FFT) and the canny edge detector for eliminating the herringbone artifact is discussed in [23]. Combined wavelet and FFT based filtering technique are introduced in [24] for vertical herringbone artifact removal. A sparse and low-rank decomposition method based on Henkel matrix is proposed in [2] for removing the motion and herringbone artifact.

The variational mode decomposition (VMD), an efficient decomposition technique decomposes the images into the number of frequency components or modes based on the characteristic center frequency [25]. Each of the frequency components has compact frequency support around the center frequency. VMD can efficiently capture all frequency variations present in the image precisely, which ensuring a fine separation of frequency components in the image. The proposed work explores the properties of VMD for separating the high-frequency non-stationary herringbone artifact from MRI. VMD is effectively used for grayscale image enhancement against Gaussian noise addition in the recently proposed works [26], [27]. The image enhancement is achieved by merely discarding the higher component modes obtained by VMD and reconstructing the enhanced image from lower component modes. However, there are chances that useful information present in the higher modes gets discarded during the enhancement. To avoid these issues, in the present work, a two-stage VMD enhancement method is proposed by further processing the higher component modes obtained from the first level of decomposition using VMD.

The paper is organized as follows: Section 2, discuss the properties of VMD, Sections 3 probes the characteristics of herringbone artifact, Section 4 presents the methodology of the proposed method, with Section 5 shows the results and observations. Conclusions are summarized in Section 6.

II. VARIATIONAL MODE DECOMPOSITION

VMD is an adaptive and fully intrinsic, variational method, which decomposes the signal into the number of sub-signals called intrinsic mode functions (IMFs) or modes. The IMFs are extracted concurrently using an iterative minimization problem. Each IMF is band limited to a center frequency which means each mode is fixed by a limited bandwidth around the characteristic center frequency [28]. The ensembles of modes can be reconstructed to the original signal optimally. The IMF is defined as the product of two functions, a slow varying amplitude modulated (AM) and a fast varying frequency modulated (FM) function. The IMF is mathematically expressed as

$$u_k(t) = A_k(t) \cos(\phi_k(t)) \quad (1)$$

The variable $u_k(t)$ represents the IMF and $A_k(t)$ is the amplitude which is a non-negative envelope and $\phi_k(t)$ is its phase. The instantaneous frequency is represented as $\omega_k(t) := \phi'_k(t)$. The bandwidth of an IMF depends on the maximum deviation and the rate of change of instantaneous frequency [29]. In order to calculate the bandwidth of the mode, a function is introduced which computes the analytic signal for each mode based on the Hilbert transform to generate a unilateral frequency spectrum [28]. The bandwidth is calculated by mixing the frequency spectrum of each mode with an exponential function, which shifts the spectrum to baseband. The constrained optimization problem is expressed as

$$\begin{aligned} \min_{u_k, \omega_k} & \left\{ \sum_k \left\| \partial_t \left[\left((\delta(t) + \frac{j}{\pi t}) * u_k(t) \right) e^{-j\omega_k t} \right] \right\|_2^2 \right. \\ & \left. s.t. \sum_k u_k = f \right. \end{aligned} \quad (2)$$

Here, $\sum_k \left\| \partial_t \left[\left((\delta(t) + \frac{j}{\pi t}) * u_k(t) \right) e^{-j\omega_k t} \right] \right\|_2^2$ represents the function for calculating the bandwidth of IMF. The constraint of this variational problem is defined as the summation of the modes, reconstruct the original signal. The variables $\{u_k\} := \{u_1, \dots, u_k\}$ represents the modes and $\{\omega_k\} := \{\omega_1, \dots, \omega_k\}$ represents the corresponding center frequencies. The minimization problem in equation 2 is expressed based on augmented Lagrangian as

$$\begin{aligned} L(\{u_k\}, \{\omega_k\}, \lambda) := & \alpha \sum_k \left\| \partial_t \left[(\delta(t) + \frac{j}{\pi t}) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \\ & \left\| f(t) - \sum_k u_k(t) \right\|_2^2 + \left\langle \lambda(t), f(t) - \sum_k u_k(t) \right\rangle \end{aligned} \quad (3)$$

This optimization formulation is solved using the iterative method called alternate direction method of multipliers (ADMM).

Similarly, 2D-VMD adaptively decompose an image into its constituent modes of spectral bands which have specific

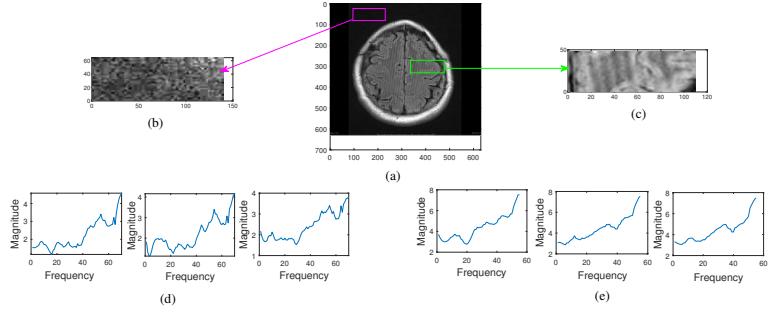


Fig. 1: Illustrates the magnitude spectrum analysis of herringbone artifact, (a) MRI corrupted with herringbone artifact, (b) noise patch, (c) image with noise patch, (d) magnitude spectra of the first three rows of pixels from the noise patch, (e) magnitude spectra of the first three rows of pixels from the image with noise patch

sparsity properties for reconstructing the original image [25]. In order to generate the 2D analytic signal, the half plane of the frequency spectrum is fixed to zero. The mathematical expression of the 2D analytic signal is represented as

$$\hat{u}_{AS,k}(\vec{\omega}) = \begin{cases} 2\vec{u}_k(\omega), & \text{if } \vec{\omega} \cdot \vec{\omega}_k > 0 \\ \hat{u}_k(\omega), & \text{if } \vec{\omega} \cdot \vec{\omega}_k = 0 \\ 0, & \text{if } \vec{\omega} \cdot \vec{\omega}_k < 0 \end{cases} \quad (4)$$

$$= (1 + \text{sgn}(\vec{\omega} \cdot \vec{\omega}_k)) \hat{u}_k(\vec{\omega})$$

The analytic signal after heterodyne demodulation and frequency mixing is expressed as

$$u_{AS,k}(\vec{x}) = u_k(\vec{x}) * \left(\delta(\langle \vec{x}, \vec{\omega}_k \rangle) + \frac{j}{\pi \langle \vec{x}, \vec{\omega}_k \rangle} \right) \delta(\langle \vec{x}, \vec{\omega}_{k,\perp} \rangle) \quad (5)$$

The constrained minimization problem for 2D-VMD is formulated as

$$\min_{u_k, \vec{\omega}_k} \left\{ \sum_k \left\| \nabla [u_{AS,k}(\vec{x}) e^{-j\langle \vec{\omega}_k, \vec{x} \rangle}] \right\|_2^2 \right\} \quad (6)$$

$$\text{s.t. } \sum_k u_k = f$$

Similar to 1D-VMD, the solution has obtained using the ADMM [25].

III. CHARACTERISTICS OF HERRINGBONE ARTIFACT

In order to obtain the spectral characteristics of the herringbone artifact in real-time MRI, we conducted a frequency domain analysis based on fast Fourier transform (FFT). An MRI corrupted with herringbone artifact is shown in Fig 1a. Two regions are observed in the MRI, a foreground region that contains the image with noise and a background region contains only noise. The most obvious way to obtain the spectral characteristics of herringbone artifact is selecting patches from both regions of MRI, where noise is common in both the patches and perform FFT on consecutive rows of pixels in each patch. The similar characteristics observed from the frequency analysis of both patches can be directly referred to as the characteristics of the herringbone artifacts. Fig 1b and 1c are the two image patches selected from background and foreground regions respectively and performed FFT on each

row of pixels. The magnitude spectra of three consecutive rows of pixels from noise patch are illustrated in Fig 1d. By visual inspection of Fig 1d, it can be observed that the magnitude spectra of each row of pixels characterized an increasing high-frequency trend. Similarly, in Fig 1e, it is evident that the magnitude spectra of three consecutive rows of pixels from the image with noise patch show an increasing high-frequency trend. The observed similar characteristics in Fig 1d and Fig 1e shows the characteristics of the herringbone artifact. The localized spectral information such as variations in magnitude and frequency are analyzed based on the short space spectrum [30]. Fig 2 represents the short space spectra of selected patches (both noise patch and image with noise patch), where Fig 2a is the short space spectrum of noise patch. The color variation observed in Fig 2 corresponds to the variation in magnitude and different level of the peaks corresponds to the variation in frequency. A similar trend can be observed from the short space spectrum of the image with noise patch, as shown in Fig 2b. The variations present in the short space spectra of both patches represent the non-stationary characteristics. In short, the non-stationary characteristics are observed in the short space spectrum of both patches due to the presence of herringbone artifact. Hence, it is essential to devise a method which efficiently removes the non-stationary noise characteristics for the effective image enhancement in the MRI.

IV. PROPOSED METHOD

In this paper, we propose an edge preserved image denoising algorithm for eliminating the non-stationary herringbone artifact from the MRI. The proposed algorithm has the following stages:

- 1) Decompose the noise-corrupted MRI using VMD into different frequency components. The modes capture all precise frequency variations present in the image. Identify and eliminate the high-frequency non-stationary components from the MRI.
- 2) The eliminated high-frequency modes are considered for another level of decomposition to extract the edges.

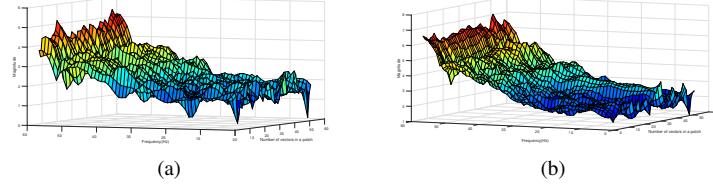


Fig. 2: Illustrates the short space spectrum analysis, (a) short space spectrum of noise patch, (b) short space spectrum of image with noise patch

The denoised image is reconstructed using low-frequency components from both levels of decomposition.

In the first stage of the proposed method, VMD decomposes the MRI into its corresponding frequency components which contain all frequency variations present in that image. The denoising is performed by eliminating the high-frequency components from the MRI. However, the eliminated high-frequency modes contain information regarding the edges. Hence, a true edge preserved image enhancement cannot be achieved using the first stage of the algorithm alone. Edges are the basic significant feature of the image. Therefore, to preserve edge information, the discarded higher modes in the first stage are further subjected to VMD decomposition in the second stage. Thus, such second level VMD based decomposition captures additional edge details in the MRI which possibly have missed in the first level processing. The enhanced MRI is reconstructed from the lower VMD modes obtained from the two stages. The block diagram of the proposed method is depicted in Fig 3. Fig 4 and 5 show the VMD decomposed modes of noisy MRI in two stages, in stage one and stage two respectively. Fig 4b is the lower mode obtained from the first stage of the proposed VMD based image enhancement. Fig 4e shows the image contains the fine details predominantly related to edge details of the MRI which is discarded from the first stage. Using the details obtained from the lower modes from the first and second stage, Fig 6c shows the reconstructed image enhanced against the herringbone noise. By comparing the noisy original image, an improvement in the perceptual quality is observed in the enhanced image. The performance of the proposed enhancement method is tested in input MRI corrupted with herringbone artifact are shown in Fig 7. Fig 8 and Fig 9 shows the results of the proposed methods in stage one and stage two respectively.

V. EXPERIMENTAL RESULTS

The dataset of MRI corrupted with herringbone artifact are collected from [31], [32]. The improved quality of the proposed method is determined based on no-reference (NR) or objective blind image quality assessment such as Blind Referenceless Image Spatial Quality Evaluator (BRISQUE) [33] and Naturalness Image Quality Evaluator (NIQE) [34]. The BRISQUE value is determined by calculating the statistics of neighboring luminance values in the spatial domain by its pairwise products. The obtained luminance coefficients

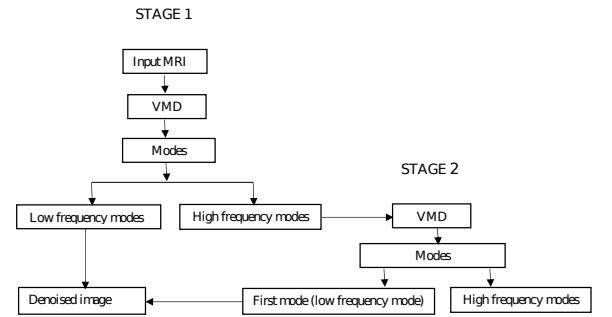


Fig. 3: Illustrates the block diagram of the proposed method

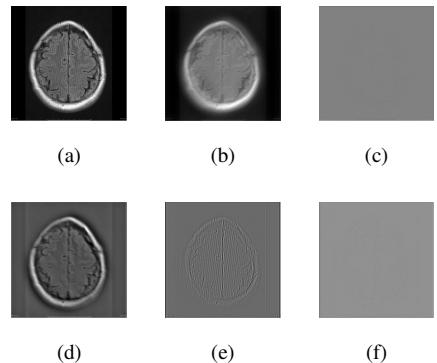


Fig. 4: Illustrates first stage decomposition of the proposed method, (a) input noisy image, (b) mode 1, (c) mode 2, (d) mode 3, (e) mode 4, (f) mode 5

are locally normalized, which quantify the naturalness of the image and quality of the image in the presence of distortion [33]. The NIQE extracts perceptually relevant features such as natural scene statistics (NSS) from local patches of image and quality-aware (QA) from the image corpus. Improved image quality is determined based on the distance between the multivariate Gaussian (MVG) fit of the NSS features and QA features [34]. The BRISQUE and NIQE score is a scalar value in the range of 0 to 100. The lowest score reflects the better perceptual quality. The results of the qualitative analysis based on BRISQUE and NIQE is given in Table I. From Table I, we can observe that the BRISQUE value in the first stage is reduced to a small value which guarantees a better

TABLE I: Performance comparison of proposed method based on BRISQUE and NIQE values

	Input Images		First stage		Second stage	
	BRISQUE	NIQE	BRISQUE	NIQE	BRISQUE	NIQE
Image 1	49.16	3.48	14.73	5.91	12.30	5.85
Image 2	41.08	5.94	21.92	3.99	20.14	3.74
Image 3	41.87	6.61	25.50	6.12	21.66	6.00
Image 4	41.73	6.83	27.82	5.82	24.99	5.73
Image 5	42.93	6.88	35.52	7.07	32.32	7.70
Image 6	45.02	6.69	34.66	4.77	21.70	4.95

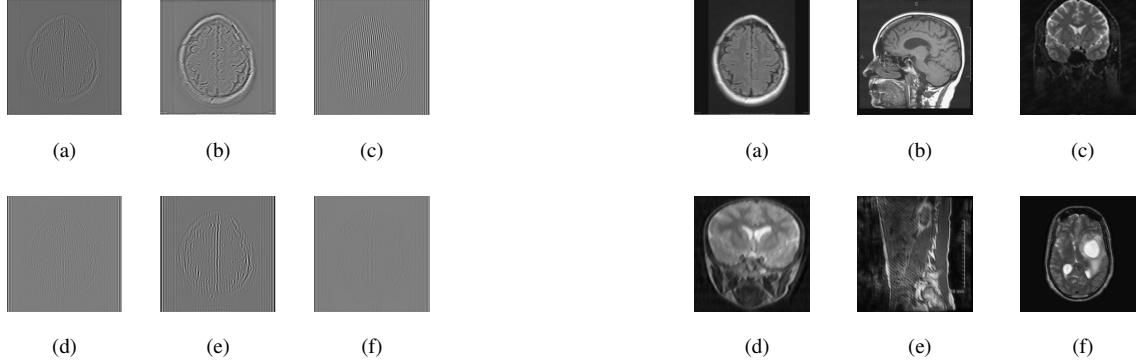


Fig. 5: Illustrates second stage decomposition of the proposed method, (a) eliminated mode from first stage, (b) mode 1, (c) mode 2, (d) mode 3, (e) mode 4, (f) mode 5

Fig. 8: Illustrates the results of first stage of the proposed method



Fig. 6: Illustrates the result of proposed method in two stages, (a) herringbone noisy image, (b) first stage output image, (c) second stage output image

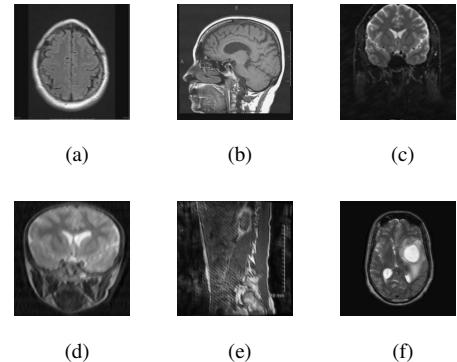


Fig. 9: Illustrates the results second stage of the proposed method

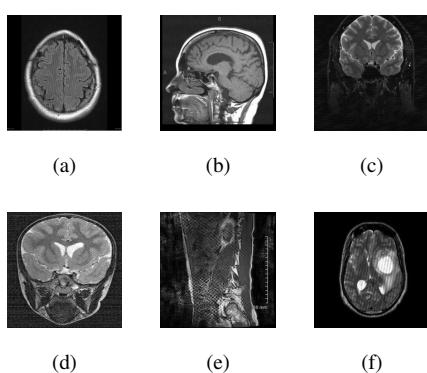


Fig. 7: Illustrates the input images corrupted with herringbone artifact

visual quality. Similarly, the NIQE value got reduced but for

certain cases, the values are increased. At the second stage, the BRISQUE and NIQE value of each image is again reduced to

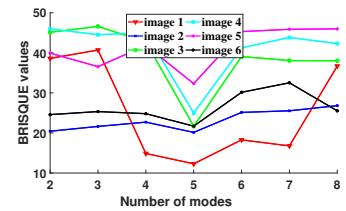


Fig. 10: Illustrates the fixing of number of modes of VMD in proposed method, the lowest BRISQUE score is obtained in mode 5 for all images.

a small value, which confirms that the edges of the image are restored. Based on the obtained quality measures, we inferred that the proposed method eliminated the herringbone artifact from the MRI and preserved its edges efficiently. In this experiment, we empirically set the number of modes in VMD as 5. Fixing the number of modes is determined based on the BRISQUE value and graphically shown in Fig 10. Variations in BRISQUE values by varying the VMD modes from 2 to 8 are examined. From the empirical experimental studies provided in Fig 10, the number modes used for the proposed algorithm is selected as 5 which provided lowest BRISQUE score.

VI. CONCLUSION

The characteristic of the herringbone artifact was studied based on the frequency analysis and found to be non-stationary. The proposed algorithm exploited the properties of VMD for eliminating non-stationary high-frequency herringbone artifact from MRI. In the two-stage algorithm, low-frequency extraction is performed in its first stage using VMD decomposition and applied the edge extraction in the second stage. The improved BRISQUE and NIQE values confirmed that the proposed algorithm removed herringbone artifact satisfactorily from the MRI by successfully preserving the edges.

REFERENCES

- [1] J. Zhuo and R. P. Gullapalli, "Aapm/rsna physics tutorial for residents mr artifacts, safety, and quality control," *RadioGraphic*, vol. 26, p. 1, 2006.
- [2] K. H. Jin, J. Lee, D. Lee, and J. C. Ye, "Sparse and low-rank decomposition of mr artifact images using annihilating filter-based hankel matrix," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 1388–1391.
- [3] K. Krupa and M. Bekiesińska-Figatowska, "Artifacts in magnetic resonance imaging," *Polish journal of radiology*, vol. 80, p. 93, 2015.
- [4] A. Stadler, W. Schima, A. Ba-Salamah, J. Kettenbach, and E. Eisenhuber, "Artifacts in body mr imaging: their appearance and how to eliminate them," *European radiology*, vol. 17, no. 5, pp. 1242–1255, 2007.
- [5] L. Erasmus, D. Hurter, M. Naudé, H. Kritzinger, and S. Acho, "A short overview of mri artefacts," *SA Journal of Radiology*, vol. 8, no. 2, pp. 13–17, 2004.
- [6] P. S. Sharma and J. N. Oshinski, "The appearance and origin of common magnetic resonance imaging artifacts, and solutions for alleviating their effects," *Medical Physics International Journal*, vol. 5, no. 1, 2017.
- [7] R. Javan, J. R. O'Rear, and J. E. Machin, "Fundamentals behind the 10 most common magnetic resonance imaging artifacts with correction strategies and 10 high-yield points." European Congress of Radiology, 2011.
- [8] S. Tirunagari, N. Poh, K. Wells, M. Bober, I. Gorden, and D. Windridge, "Movement correction in dee-mri through windowed and reconstruction dynamic mode decomposition," *Machine Vision and Applications*, vol. 28, no. 3-4, pp. 393–407, 2017.
- [9] B. Lorch, G. Vaillant, C. Baumgartner, W. Bai, D. Rueckert, and A. Maier, "Automated detection of motion artefacts in mr imaging using decision forests," *Journal of medical engineering*, vol. 2017, 2017.
- [10] W. Lin, F. W. Wehrli, and H. K. Song, "Correcting bulk in-plane motion artifacts in mri using the point spread function," *IEEE transactions on medical imaging*, vol. 24, no. 9, pp. 1170–1176, 2005.
- [11] J. Mohan, V. Krishnaveni, and Y. Guo, "A survey on the magnetic resonance image denoising methods," *Biomedical Signal Processing and Control*, vol. 9, pp. 56–69, 2014.
- [12] R. M. Henkelman, "Measurement of signal intensities in the presence of noise in mr images," *Medical physics*, vol. 12, no. 2, pp. 232–233, 1985.
- [13] E. R. McVeigh, R. M. Henkelman, and M. J. Bronskill, "Noise and filtration in magnetic resonance imaging," *Medical physics*, vol. 12, no. 5, pp. 586–591, 1985.
- [14] J. Mohan, V. Krishnaveni, and Y. Guo, "Mri denoising using nonlocal neutrosophic set approach of wiener filtering," *Biomedical Signal Processing and Control*, vol. 8, no. 6, pp. 779–791, 2013.
- [15] K. Krissian and S. Aja-Fernández, "Noise-driven anisotropic diffusion filtering of mri," *IEEE transactions on image processing*, vol. 18, no. 10, pp. 2265–2274, 2009.
- [16] A. Buades, B. Coll, and J.-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 490–530, 2005.
- [17] P. Coupé, P. Yger, S. Prima, P. Hellier, C. Kervrann, and C. Barillot, "An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images," *IEEE transactions on medical imaging*, vol. 27, no. 4, pp. 425–441, 2008.
- [18] J. V. Manjón, P. Coupé, A. Buades, D. L. Collins, and M. Robles, "New methods for mri denoising based on sparseness and self-similarity," *Medical image analysis*, vol. 16, no. 1, pp. 18–27, 2012.
- [19] T. K. Foo, N. S. Grigsby, J. D. Mitchell, and B. E. Slayman, "Snore: spike noise removal and detection," *IEEE transactions on medical imaging*, vol. 13, no. 1, pp. 133–136, 1994.
- [20] Y.-H. Kao and J. R. MacFall, "Correction of mr k-space data corrupted by spike noise," *IEEE transactions on medical imaging*, vol. 19, no. 7, pp. 671–680, 2000.
- [21] X. Zhang, P.-F. Van De Moortele, J. Pfeuffer, and X. Hu, "Elimination of k-space spikes in fmri data," *Magnetic resonance imaging*, vol. 19, no. 7, pp. 1037–1041, 2001.
- [22] A. E. Campbell-Washburn, D. Atkinson, Z. Nagy, R. W. Chan, O. Josephs, M. F. Lythgoe, R. J. Ordidge, and D. L. Thomas, "Using the robust principal component analysis algorithm to remove rf spike artifacts from mr images," *Magnetic resonance in medicine*, vol. 75, no. 6, pp. 2517–2525, 2016.
- [23] T. D. Vishnumurthy, H. S. Mohana, V. A. Meshram, and P. Kammar, "A novel algorithm for removal of herringbone artifact in brain mr images using fft and canny edge detector," *American Journal of Engineering Research*, vol. 5, pp. 184–189, 2016.
- [24] ———, "Suppression of herringbone artifact in mr images of brain using combined wavelet and fft based filtering technique," *International Journal of Computer Sciences and Engineering*, vol. 4, no. 2, pp. 66–71, 2016.
- [25] K. Dragomiretskiy and D. Zosso, "Two-dimensional variational mode decomposition," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2015, pp. 197–208.
- [26] S. Lahmiri and M. Boukadoum, "Biomedical image denoising using variational mode decomposition," in *Biomedical Circuits and Systems Conference (BioCAS), 2014 IEEE*. IEEE, 2014, pp. 340–343.
- [27] G. S. C. Kumar, R. K. Kumar, G. A. Naidu, and J. Harikiran, "Noise removal in microarray images using variational mode decomposition technique," *Telkomnika*, vol. 15, no. 4, 2017.
- [28] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE transactions on signal processing*, vol. 62, no. 3, pp. 531–544, 2014.
- [29] C. Aneesh, S. Kumar, P. Hisham, and K. Soman, "Performance comparison of variational mode decomposition over empirical wavelet transform for the classification of power quality disturbances using support vector machine," *Procedia Computer Science*, vol. 46, pp. 372–380, 2015.
- [30] B. Hinman, J. Bernstein, and D. Staelin, "Short-space fourier transform image processing," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, vol. 9. IEEE, 1984, pp. 166–169.
- [31] MRI image data set with herringbone noise (source: radiopaedia.org/articles/herringbone-artifact).
- [32] MRI image data set with herringbone noise (source: mriquestions.com/data-artifacts.html).
- [33] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [34] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

A Weighted SVM Based Approach for Automatic Detection of Posterior Myocardial Infarction Using VCG Signals

Eedara Prabhakararao and Samarendra Dandapat

Department of Electrical and Electronics Engineering

Indian Institute of Technology Guwahati

Guwahati-781039, Assam, India

Email: {e.prabha, samaren}@iitg.ac.in

Abstract—Myocardial infarction (MI), commonly known as heart attack is a life-threatening arrhythmia occurs due to insufficient oxygen supply to the heart tissues resulted from formation of clots in one or more coronary arteries. There is a growing interest among researchers for automatic detection of MI using computer algorithms. Based on the spatial location of damaged tissues MI is further categorized as anterior MI, septal MI, lateral MI, inferior MI and posterior MI. Among all, automatic detection of posterior MI (PMI) with standard 12-lead electrocardiogram (12-lead ECG) signal is challenging as it does not have monitoring electrodes posterior to human body. In this paper, we propose an automatic method for PMI detection using 3-lead vectorcardiogram (3-lead VCG) signal. The proposed approach exploits changes in electrical conduction properties of heart tissues during cardiac activity for healthy control (HC) and PMI subjects in three-dimensional (3D) space. To quantify these changes multiscale eigen features (MSEF) of subband matrices are used. Furthermore, we propose a cost sensitive weighted support vector machine (WSVM) classifier to combat class imbalance, which is a common problem in real-world disease data classification. The publicly available PhysioNet/PTBDB diagnostic database has been used to validate the proposed method by using a total of 1463 HC, and 148 PMI 4 sec 3-lead VCG signals. The best test accuracy of 96.69%, sensitivity of 80%, and geometric mean of 88.72% are achieved by WSVM classifier with radial basis function (RBF) kernel.

Keywords—12-lead ECG, VCG, conduction properties, posterior MI, multiscale features, data imbalance, weighted SVM

I. INTRODUCTION

Myocardial infarction (MI), commonly occurs due to insufficient oxygen supply to the heart tissues manifest from formation of clots in coronary arteries [1]. It can takes place in different portions of the heart including, inferior, anterior, septal, posterior, lateral, inferior-lateral, septal-anterior, and posterior-lateral [1], [2]. MI damages heart myocardial tissues due to insufficient supply of oxygen and nutrients. This damage turns the myocardial tissues into either slow conductor or complete insulator depending on the severity of the occlusion and, consequently leads to variation in the normal electrical conduction through the heart tissues (i.e., depolarization and repolarization) [2].

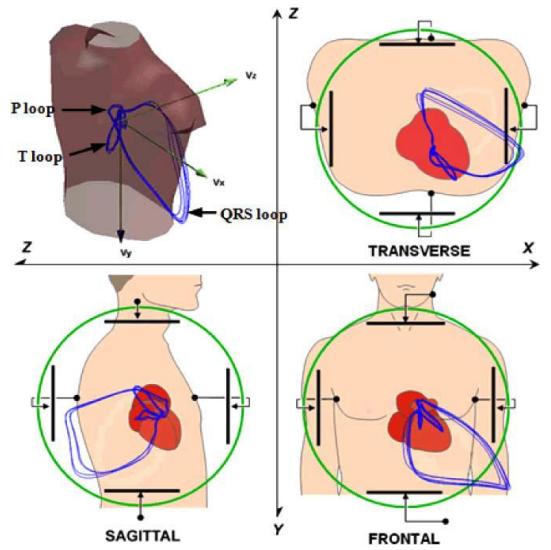


Fig. 1. Depicts different monitoring views of human torso showing the position of electrodes for VCG signal acquisition and vector loops in transverse, sagittal, and frontal planes along with corresponding coordinate axis [11].

The gold standard 12-lead ECG (12-lead ECG) is often often preferred by cardiologists for routine cardiac monitoring and arrhythmia detection [1]. However, due to unavailability of direct measurements from the posterior wall of the heart, the 12-lead ECG reported to be insensitive for diagnosing posterior ST elevation MI (PMI) [3] - [9]. In clinical practice, ST depression in the anterior leads (V1 to V4) are often preferred for PMI diagnosis. However, ST depression in anterior leads is neither specific nor sensitive for PMI detection as these changes have similar diagnostic characteristics as the anterior ischemia. This similarity makes PMI diagnosis challenging and, even difficult for the cardiologists to inspect [4], [6].

The posterior leads (V7, V8, and V9), placed posterior to the human torso towards left side are the most preferred leads for PMI diagnosis [5], [7]. But, these leads are not commonly available in the gold standard 12-lead ECG systems, since, it makes cardiac monitoring systems complex with 13

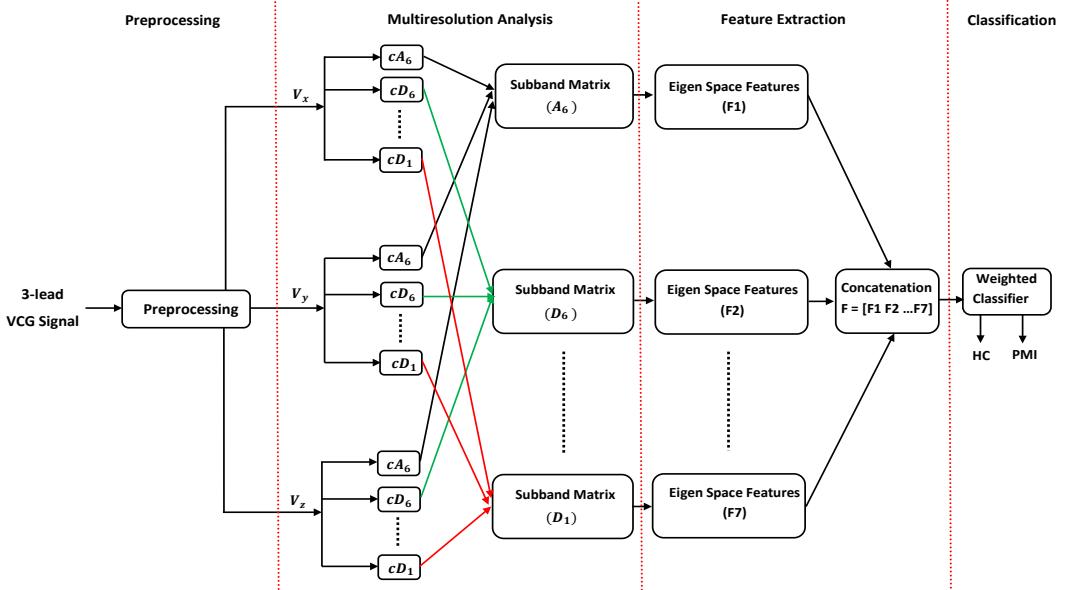


Fig. 2. Flow diagram of proposed four stage algorithm for posterior myocardial infarction detection.

electrodes for recording 15-lead ECG and it also hampers patient comfort. Therefore, VCG or Frank XYZ system is an alternative multilead ECG recording system with less number of electrodes (seven), more interpretable and helpful in understanding the dipole hypothesis of heart electrical system [10], [11]. Fig. 1 shows that the VCG signals (V_x , V_y , and V_z) capture the heart electrical activity in 3 orthogonal directions (X,Y,Z) and can be projected into different 2D planes including transverse, sagittal and frontal [2]. The cardiac activity of P, QRS, and T-wave can be seen as the near periodic patterns of loops for each cardiac cycle as illustrated in Fig.1.

To the best of our knowledge there is no article in the literature reporting PMI diagnosis using VCG signals and handling class imbalance problem during classification. However, there are few works reported for PMI detection using multi-channel ECG signals (12-lead or 15-lead). These methods use amplitude, duration and ST elevation or ST depression features to diagnose PMI [5], [6], [9]. Zhou *et al.* [5], Aqel *et al.* [9] reported that, a minimum improvement in sensitivity of 30% is observed while using 15-lead ECG over 12-lead ECG. Similarly, Din *et al.* [6], shows the inability of 12-lead ECG for PMI diagnosis with very low accuracy of 55%.

In the aforesaid methods, PMI diagnosis is done by using morphological features of the multi-lead ECG signals. However, it is difficult to extract these features reliably during ECG signals corrupted with different noises [12]. In [4], Gorselen *et al.* even used a cardiologist for visual inspection of PMI events. However, manual inspection of long duration multi-lead ECG signals for identifying PMI events is impractical [12]. To date, to the best of our knowledge, there is no PMI detection algorithm that takes into consideration of the data imbalance problem during classification. In real-world, HC subjects are largely available in comparison to the PMI

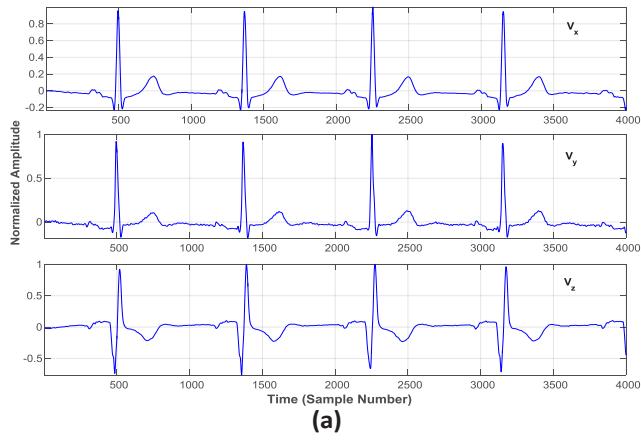
patients. It is already reported that, 12-lead ECG signals are inadequate for reliable automatic diagnosis of PMI.

Therefore, in this paper we propose a novel algorithm for automatic detection and classification of PMI from HC subjects using 3-lead VCG signals. The proposed method consists of 4 stages as shown in Fig. 2. In the first stage 3-lead VCG signals are preprocessed to remove baseline wander (BW) noise using a butterworth low pass filter (LPF) and high frequency (HF) noise using a 3-point moving average (MA) filter. In second stage, multiscale subband matrices (MSSM) are constructed from a preprocessed VCG signal using a 6 level wavelet decomposition. The third stage exploits covariance structures of selected MSSM to obtain a 12-dimensional multiscale eigen features (MSEF) for efficient classification of HC and PMI using a weighted support vector machine (WSVM) classifier.

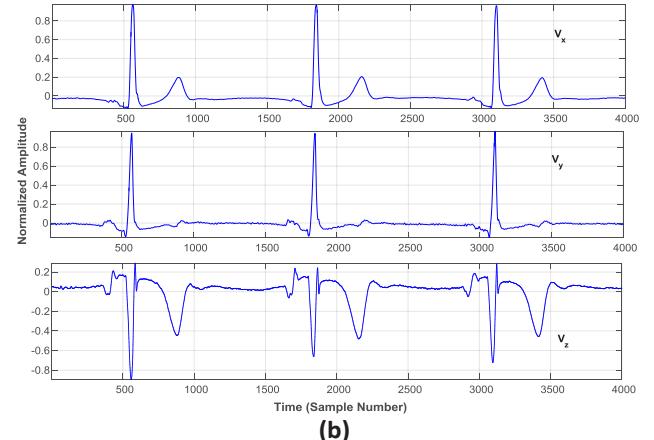
II. PROPOSED METHOD AND MATERIAL

A. Database Selection and Preprocessing

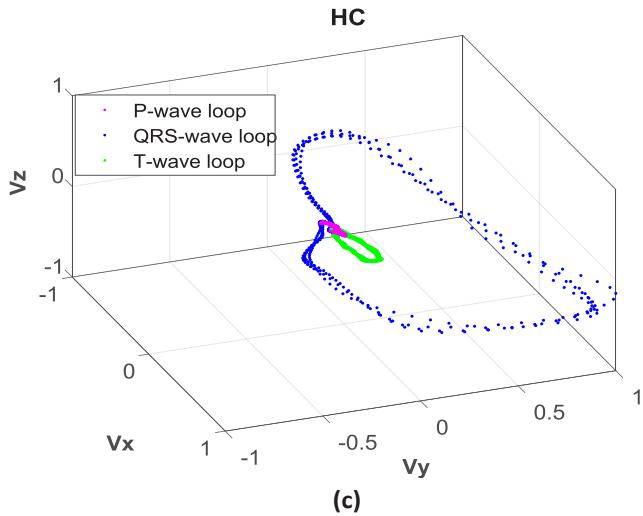
The proposed method is evaluated using the 3-lead VCG records obtained from the publicly available PhysioNet/PTBDB diagnostic database [13]. This database consists of a total of 549 records from 290 subjects with varying cardiac abnormalities. Each record consisting of 15 simultaneously recorded lead signals i.e., 12-lead ECG (lead I, lead II, lead III, aVR, aVF, aVL, V1, V2, V3, V4, V5, V6) and 3-lead VCG (V_x , V_y , V_z) signals. Each lead signal is digitized by sampling it at 1 kHz with 16-bit resolution. In this study, from the available 52 HC subjects a total of 1463 segments of 4 sec duration VCG signals are selected. Similarly, a total of 148 segments are obtained from 4 PMI subjects. The selected dataset contains more HC segments than the PMI. This leads to a imbalance in the dataset, which may bias the learning



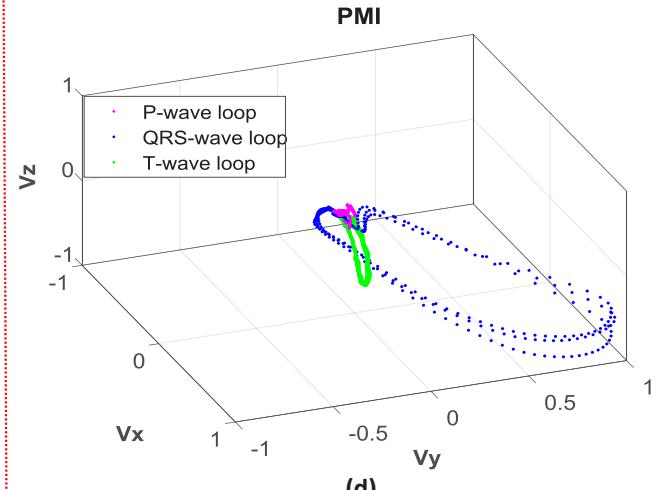
(a)



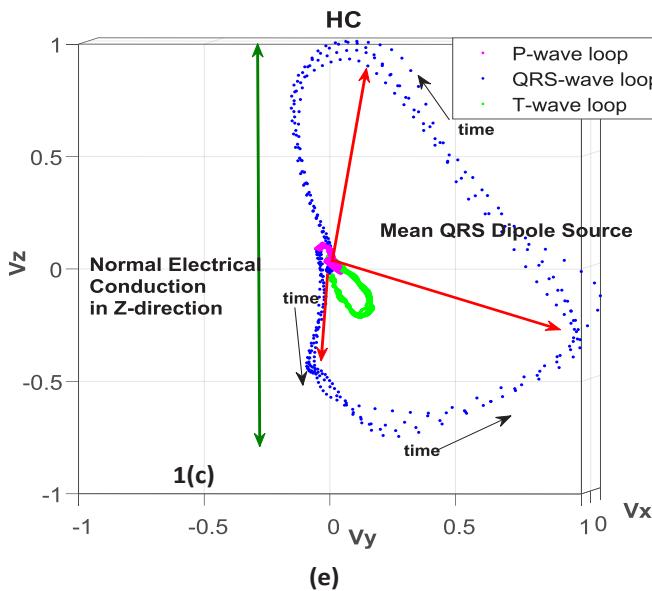
(b)



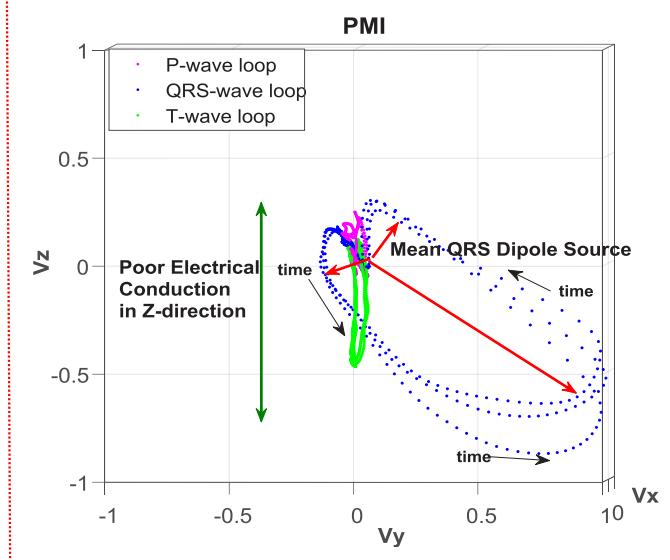
(c)



(d)



(e)



(f)

Fig. 3. Depicts difference in electrical conduction for HC and PMI 3-lead VCG signals taken from PTBDB database, (a)-(d) illustrates HC and PMI 3-lead VCG signals, its 3D vector loops of P, QRS, T-waves , (e) and (f) illustrates 2D sagittal view of heart electrical activity for the HC and PMI subjects respectively.

algorithm. From the selected segments, BW noise is removed by a butterworth LPF with cut-off frequency 0.679 Hz [12], and HF noise is reduced by a 3-point MA filter. Later, these preprocessed segments are used for multiscale analysis to extract 12-dimensional MSEF for efficient classification of HC and PMI subjects.

B. Multiscale Feature Extraction

The preprocessed VCG signals are analysed in the wavelet domain to extract discriminative eigen features for the classification. PMI causes variation in electrical conduction properties of heart tissues, which leads to change in the morphological features of VCG signals can be seen in Fig.3. Fig. 3(a)-(d) show the 3-lead VCG signals and corresponding 3D contours of P, QRS, and T-wave of HC and PMI subjects. Fig. 3(e) and (f) show the saggital view of these contours in 2D plane. It can be seen that, there is a significant difference in the electrical conduction along the Z-direction. In order to quantify these changes, covariance structures of the MSSM are analysed. The detailed discription of the proposed MSEF are discussed in the following subsections.

1) Discrete Wavelet Transform of 3-lead VCG: Wavelet transform has the ability to represent any finite energy non-stationary signal, grossly in to different subbands [14]. In this paper, we used wavelet transform to decompose each lead of 3-lead ECG to capture low frequency (LF) local waves (P-wave, T-wave), low and high frequency (HF) QRS-complex information, and other HF artefact in different subbands [15]. The dyadic wavelet transform uses a multiresolution pyramidal decomposition technique based filter bank implementation for $R + 1$ subbands decomposition with R as level of decomposition. For the l^{th} VCG lead with k^{th} wavelet coefficient, the decomposition results an approximation subband coefficients $cA_{R,k}^l$ at level R , and at detail subbands $cA_{r,k}^l$, at level r with $r = 1, 2, \dots, R$. The approximation and detail coefficients are derived from the inner or dot product of each lead of 3-lead VCG signal with scaling function $\phi_{R,k}[n]$, and wavelet function $\psi_{r,k}[n]$ given by (1), (2).

$$\phi_{R,k}[n] = 2^{-R/2} \phi(2^{-R} n - k) \quad (1)$$

$$\psi_{r,k}[n] = 2^{-r/2} \psi(2^{-r} n - k) \quad (2)$$

The approximation and detailed subband coefficients of 3-lead VCG $x^l[n]$ are calculated as $cA_{R,k}^l = \langle x^l[n], \phi_{R,k}[n] \rangle$ and $cD_{r,k}^l = \langle x^l[n], \psi_{r,k}[n] \rangle$ [15]. In this paper, for a sampling rate of 1 kHz, and *bior6.8* mother wavelet [15], a 6 level wavelet decomposition is employed to capture diagnostic information of VCG signal distributed over different wavelet subbands based on their frequency content. It has been seen from many studies that, most of the clinically relevant information of the ECG signal can be obtained from lower frequency subbands [15]. Since, all 3-lead VCG signals are decomposed with same mother wavelet and decomposition levels, results in similar subbands with an same number of coefficients. By using this structure of 3-lead VCG subbands, $R + 1$ MSSM can be formed by arranging similar subbands

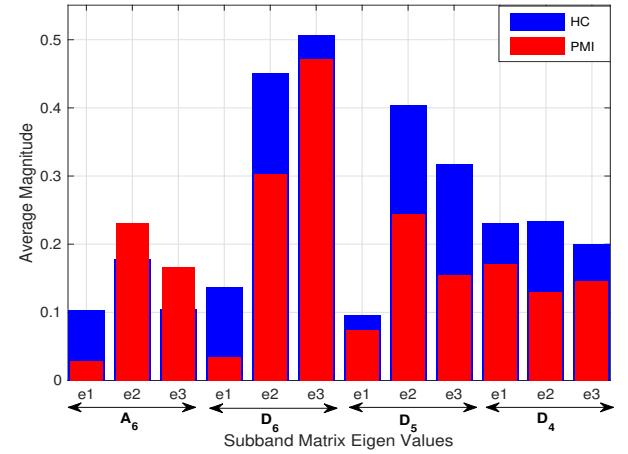


Fig. 4. Illustrates average magnitude of eigen values for 140 HC and PMI subjects across selected subband matrices (A_6 , D_6 , D_5 , and D_4)

of all leads into one subband matrix as given by (3), (4), and illustrated in Fig.2,

$$A_R = [cA_{R,k}^1, cA_{R,k}^2, \dots, cA_{R,k}^l] \quad (3)$$

$$D_r = [cD_{r,k}^1, cD_{r,k}^2, \dots, cD_{r,k}^l], \quad (4)$$

where $l = 3$ is the number of leads of 3-lead VCG.

2) Multiscale Subband Eigen Features: As we discussed earlier, the subband matrices A_R and D_r contain similar diagnostic information from different leads. The covariance among the leads can be more reliably quantified using eigen space features of subband matrices. The covariance matrices (CM) from mean removed data are calculated as,

$$C_{A_R} = \frac{1}{N_R - 1} ([A_R]^T [A_R]) \quad (5)$$

$$C_{D_r} = \frac{1}{N_r - 1} ([D_r]^T [D_r]) \quad (6)$$

where N_R and N_r are number of coefficients in approximation and detail subbands and C_{A_R} represents approximation CM at level R and C_{D_r} represents detail CM at level r . In this study, we extracted MSEF from $C_{A_6}, C_{D_6}, C_{D_5}, C_{D_4}$ subband CM of each 3-lead VCG as most of the diagnostic information contained in those subbands [15]. As shown in Fig. 2, a 12-dimensional feature vector is formed by concatenating features from each MSSB. The orthogonal eigen vectors give the directions of maximum variances of the data and eigen values gives the magnitude of variance [16]. It is observed that, variation of the data for example in Z-direction is more for HC than PMI subjects. Therefore, the average eigen values should have more magnitude for HC than PMI. It can be seen from Fig. 4 that, the average magnitude of eigen values from selected subband matrices of 140 HC and 140 PMI subjects have clear discrimination, especially, for the QRS-wave dominant D_5, D_4 MSSM. Therefore, those features are fed to the proposed WSVM classifier for automatic HC and PMI classification.

TABLE I
CLASSIFICATION RESULTS OF KNN, SVM AND PROPOSED WEIGHTED SVM CLASSIFIERS

Classifier	TP	FN	TN	FP	Accuracy(%)	Specificity(%)	Sensitivity(%)	Gmean(%)
KNN	13	32	428	11	91.11	97.49	28.88	53.06
SVM-linear	24	21	434	5	94.62	98.86	53.33	72.61
Weighted SVM-linear	29	16	431	8	95.04	98.17	64.44	79.54
SVM-RBF	26	19	439	0	96.07	100	57.77	76.01
Weighted SVM-RBF	36	9	432	7	96.69	98.40	80	88.72

C. Classifiers and Evaluation Measures

In this paper, we explored three supervised binary classifiers, k-nearest neighbours (KNN), SVM and weighted SVM (with linear and non-linear kernel) classifiers for labelling any test feature vector from 3-lead VCG to either HC or PMI [17]. For KNN with euclidean distance as a measure, a total of 5 nearest neighbours are used for the classification. The SVM is a popular binary supervised classifier originally proposed by Vapnik [19].

Assume that a dataset is represented by a set $D = \{(x_i, y_i)\}_{i=1}^N$, where $(x_i, y_i) \in \mathcal{R}^{d+1}$, N and d are the number of samples and features respectively. Each x_i represents a sample with n features and $y_i \in \{1, -1\}$ represents class label. The SVM classifies data points by finding a separating hyperplane, whose distance is maximum from data samples of each class. The optimal separating hyperplane can be defined by w and b parameters can be obtained by solving the convex optimization problem [17]. The general soft margin SVM with related quadratic programming problem can be defined as,

$$\min. \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i \quad (7)$$

$$s.t. \quad y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \quad i = 1, \dots, N \quad (8)$$

where $\xi_i \geq 0$ for $i = 1, \dots, N$ and ϕ is the kernel function. The parameter C denotes cost for misclassification or penalization. However, with this formulation each data sample gets equal importance in the learning process. This may not be desirable when one class contains very less samples than other as learning algorithms may tend to bias [19]. This data imbalance is very common problem in many real-world applications including disease classification using medical data, behavior analysis, fraud detection in banking operations, sentiment analysis etc., [18]. Therefore, to handle this problem a cost sensitive modified soft margin SVM has been proposed to make learning process less bias towards the majority class. In this modified SVM, we use different penalty or weight for positive class (C^+) and negative class (C^-) and those weights are inversely proportional to number of samples in each class. The modified problem can be formulated as,

$$\min. \frac{1}{2} \|w\|^2 + \left(C^+ \sum_{\{i|y_i=+1\}}^{n^+} \xi_i \right) + \left(C^- \sum_{\{j|y_j=-1\}}^{n^-} \xi_j \right) \quad (9)$$

$$s.t. \quad y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \quad i = 1, \dots, N \quad (10)$$

where $\xi_i \geq 0$ for $i = 1, \dots, N$, n^+ and n^- are number samples in positive and negative class respectively. In this paper, we used Gaussian non-linear kernel or radial basis function $k(x_t, x_n)$ with x_t as test feature, x_n as support vectors expressed in the feature space for RBF-SVM and RBF-WSVM classifiers. The RBF kernel has variance parameter γ shown in (11), to control flexibility of resulting classifier usually obtained from cross validation process.

$$k(x_t, x_n) = \exp(-\gamma \|x_t - x_n\|^2) \quad (11)$$

The performance of proposed binary classification method is evaluated using four standard benchmark statistical measures, with sensitivity se ; measures the portion of PMI subjects being classified as PMI, and specificity sp ; measures the portion of HC subjects being classified as HC, accuracy acc ; portion of the subjects that have been classified correctly, and Geometric mean Gm ; often used measure to validate imbalanced datasets [12]. These measures are quantified by using number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as, $se(\%) = \left(\frac{TP}{TP + FN} \right) \times 100$, $sp(\%) = \left(\frac{TN}{TN + FP} \right) \times 100$, $acc(\%) = \left(\frac{TP + TN}{TP + TN + FP + FN} \right) \times 100$, and $Gm(\%) = \sqrt{(se(\%) \times sp(\%))}$.

III. RESULTS AND DISCUSSION

In this section, the evaluation of proposed PMI detection algorithm with suitable measures has been discussed. In the first stage of proposed method, a total of 1463 3-lead VCG HC segments and a total of 148 3-lead VCG segments are preprocessed and a 6 level multiscale subbands are decomposed, later 12-dimensional eigen features are evaluated for selected subband matrices as discussed in previous section. Here, each segment duration chosen to be 4 sec as cardiac alarm should be triggered within 10 sec after an arrhythmia detected [12]. In this study, a total of 1024 HC, and 103 PMI samples are used for training, and a total of 439 HC, and 45 PMI samples are used for testing. Finally, training and testing sets are passed to KNN, SVM, weighted SVM classifiers for further evaluation. The balanced dataset (HC -148, PMI - 148) with RBF kernel achieves overall accuracy of 61.46%, and positive class sensitivity of 52.08% which is poor in-terms of disease diagnosis. In this paper, we combat this by using all HC samples and available PMI samples through weighted SVM classifier. The weights for WSVM are calculated as

$C^- = C(\frac{N}{n^-}) = (1.42)C$, $C^+ = C(\frac{N}{n^+}) = (9.88)C$, where N is number of data points, and RBF γ is of 0.125 obtained from a grid search cross validation procedure.

From Table I, it is observed that, even though accuracy and specificity of KNN, SVM-linear, and SVM-RBF are high and more than 91%, and 97% respectively, but these are not a reliable measures for imbalanced datasets as all points might be classified as one class (majority). However, sensitivity and Gmean measures for those classifies are very less and not reliable for PMI detection. Since, the proposed weighted SVM classifier hypothesized to handle this data imbalance problem has been supported by obtained results given in Table I. It is observed that, there is a significant improvement in the sensitivity and Gmean measures with a slight improvement in accuracy and slight reduction in specificity. The best performance on test data was obtained by weighted SVM-RBF classifier as $se(\%) = 80$, $Gm(\%) = 88.72$, and $acc(\%) = 96.69$.

TABLE II
PERFORMANCE COMARISION WITH EXISTING METHODS

Author	lead system	features	acc(%)	se(%)	Gm(%)
Zhou <i>et al.</i> , [5]	15-lead ECG	Amp,Dur ^a	-	68	-
Din <i>et al.</i> , [6]	12-lead ECG	Amp,Dur ^a	55	-	-
Aqel <i>et al.</i> , [9]	15-lead ECG	ST ele ^b	-	62	-
Proposed	3-lead VCG	12 eigen	96.69	80	88.72

^a Amp,Dur: Amplitude and duration

^b ST ele : ST elevation

From Table II, it can be observed that proposed method with 12-dimensional eigen features from only 3-lead VCG signal outperforms existing PMI detection methods with huge improvement in accuracy and sensitivity. All existing methods uses multi-channel ECG (12-lead or 15-lead) signals with unreliable or less robust features including amplitude, duration , and ST elevation or depression. Some papers reported with manual inspection from cardiologist, but those existing methods are not suitable for automatic diagnosis of PMI using long duration multi-channel ECGs with possible physiological noises. Therefore, the proposed method with 3-lead VCG is best suitable for quick diagnosis of PMI and helpful for cardiologists to improve Percutaneous Coronary Intervention (PCI) treatment for posterior wall infarction patients.

IV. CONCLUSION

In this paper, we presented a novel yet simple automatic PMI detection algorithm from HC subjects using 3-lead VCG signals. The proposed method consist of four stages, i) first stage perform VCG signal preprocessing to reduce BW and HF noises ,ii) in the second stage, multiscale subband matrices (MSSM) are constructed from preprocessed VCG signal using a 6 level wavelet decomposition, iii) The third stage exploits covariance structures of selected MSSM to obtain a 12-dimensional MSEF for efficient classification, iv) last stage performs supervised binary classification of 3-lead VCG signals as HC or PMI using weighted SVM classifier to combat data imbalance problem. The proposed method is evaluated

on PTBDB database and the best test accuracy of 96.69%, sensitivity of 80%, Gmean of 88.72% are achieved by 12-dimensional MSEF and weighted SVM-RBF classifier. The proposed method does not require complex QRS detectors, amplitude and duration features computation from 12-lead / 15-lead ECG signals. However, it uses simple multiscale subband lead-correlations of 3-lead VCG signals, and suitable for guiding cardiologists to improve PCI treatment.

REFERENCES

- [1] V. Fuster, R. Walsh and R. A. Harrington, "Hurst's the Heart: Manual of Cardiology," *The McGraw-Hill Companies*, 2013.
- [2] H. Yang, "Multiscale Recurrence Quantification Analysis of Spatial Cardiac Vectorcardiogram Signals," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 2, pp. 339-347, Feb. 2011.
- [3] PJ. Zimetbaum, and ME. Josephson, "Use of the Electrocardiogram in acute Myocardial Infarction," *New Englad Jour. Med.*, June, 2003.
- [4] E.O.F. van Gorselen, F.W.A. Verheugt, B.T.J. Meursing, A.J.M. Oude Ophuis, "Posterior myocardial infarction: the dark side of the moon," *Netherland Heart Jour.*, vol. 15, Jan, 2007.
- [5] S. Zhou *et al.*, "An automated algorithm to improve ECG detection of posterior STEMI associated with left circumflex coronary artery occlusion," *Comput Cardiol.*, pp. 3336, 2006.
- [6] I. Din, M. Adil, Hameedullah, M. Faheem, F. Abbas, and M. Hafizullah, "Accuracy of 12 lead ECG for diagnosis of posterior myocardial infarction," *Jour. Postgrad. Med. Inst.*, 2014.
- [7] J. T. Lewis, "ECG Diagnosis: Isolated Posterior Wall Myocardial Infarction," *Clinical Med. The Permanente Jour.*, vol. 19, 2015.
- [8] W. Brady, R. Harrigan, and T. Chan, "The Diagnosis: Acute Posterior Wall Myocardial Infarction," *Cases in Electrocardiography*, Sept. 2002.
- [9] R. A. Aqel *et al.*, "Usefulness of Three Posterior Chest Leads for the Detection of Posterior Wall Acute Myocardial Infarction," *American Jour. Cardio.*, pp. 159-164, 2009.
- [10] I. Tomasic and R. Trobec, "Electrocardiographic systems with reduced numbers of leads - synthesis of the 12-lead ecg," *IEEE Rev. Biomed. Eng.*, vol. 7, pp. 126-142, 2014.
- [11] J. Malmivuo and R. Plonsey, "Bioelectromagnetism : Principles and Applications of Bioelectric and Biomagnetic Fields," *1st edition Oxford Univ. Press, USA*, Jul 1995.
- [12] G. D. Clifford and G. B. Moody, "Signal quality in cardiorespiratory monitoring," *Physiol. Meas.*, vol. 33, no. 9, p. 6, 2012.
- [13] M. Oeff *et al.* (2012,) "The PTB diagnostic ECG database," *Natational Metrology Institute Germany [Online]*. Available: <http://www.physionet.org/physiobank/database/ptbdb>.
- [14] C. Orphanidou, and I. Drobnjak, "Quality assessment of ambulatory ECG using wavelet entropy of the HRV signal," *IEEE Jour. of Biomed. and Health Infor.*, vol. 21, no. 5, pp. 1216-1223, sep. 2017.
- [15] L. N. Sharma, R. K. Tripathy and S. Dandapat, "Multiscale Energy and Eigenspace Approach to Detection and Localization of Myocardial Infarction," *IEEE Tran. on Biomed. Eng.*, vol. 62, no. 7, pp. 1827-1837, July 2015.
- [16] J. Shlens, "A Tutorial on Principal Component Analysis," 2005.
- [17] C. C. Chang, and C. J. Lin, "LIBSVM-A library for support vector machines," *[Online]*. Available: <http://www.csie.ntu.edu.tw/cjlin/libsvm/>.
- [18] H. He, and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Know. Data Eng.*, vol. 21, pp. 1263-1284, 2009.
- [19] Y. Tang, Y. Zhang, N. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Systems, Man, and Cyber., Part B*, vol. 39, pp. 281-288, 2009.

Brain Tumor Segmentation Using Discriminator Loss

Joydeep Das¹, Rashmin Patel¹, and Vinod Pankajakshan²

¹Department of Electronics and Communication Engineering, IIT Roorkee

¹{joydeepdas994,rashminpatel405}@gmail.com

²{vinodfec}@iitr.ac.in

Abstract—The emerging field of Computer Vision has found enormous applications in our day-to-day lives and Medical Image Processing is one of the most prominent fields among them. Brain Tumor Segmentation is an important and challenging task because of the variety in shapes, sizes and texture content of the various types of brain tumors. Specifically, MICCAI BraTS organizes Brain Tumor Segmentation challenge every year. Since the evolution of CNNs it has obtained state-of-the-art results in the majority of computer vision related tasks. On BraTS Challenge 2017, an ensemble average of various CNN models (EMMA) holds the state-of-the-art performance. In this paper, we have proposed a model inspired by the classic Generative Adversarial Network (GAN). The proposed network has two models namely, Generator or Segmentor which generates label map of the input image and a Discriminator which helps the Generator model for an optimum solution by taking into account both short as well as long-distance spatial correlations between pixels with the help of a novel multi-scale loss function. The proposed architecture has three GANs in a cascaded fashion, each for Whole Tumor, Tumor Core and Enhancing Tumor, where the former network helps in effective reduction of false positives for the later networks. Our method also employs a multi-scale loss function derived from intermediate layers of Discriminator rather than depending just on a final layer cross-entropy loss. A multi-scale loss function also reduces unnecessary smoothing on contours. The proposed method performed comparatively better than the state-of-the-art techniques, having Dice scores of 0.820, 0.874 and 0.783 for Enhancing Tumor, Whole Tumor and Tumor Core respectively.

Index Terms—Cascaded GAN, Content and Adversarial Loss

I. INTRODUCTION

Cancer can be termed as unnatural or uncontrolled growth of the tissue cells. When inside the brain, the brain tissue grows in an irregular fashion we call it as a brain tumor. The cancerous tissues that originate from the glial cells are called Gliomas, are the most commonly found brain cancers. Gliomas can be divided into two types, first the Low Grade Glioma (LGG) which are not so dangerous and belong to the early stage of development of cancer cells and second is the High Grade Glioma (HGG) which has high mortality rates. Therapy techniques are used over these cancer tissue where care has to be taken not to destroy the healthy brain tissue. Thus to reduce this collateral damage during therapy techniques, proper identification of cancerous region against the healthy tissue is very essential.

Magnetic Resonance Imaging (MRI) scans of four different types are used to capture contrasted tissue images. The four modes used are Fluid Attenuation Inversion Recovery (FLAIR), native T1, T2-weighted (T2) and T1-weighted (T1Gd). In this task, we are required to develop a method which assigns pixel wise labelling to different sub regions of gliomas namely Tumor Core (TC), Whole Tumor (WT) and Enhancing Tumor (ET) with the help of a given clinically-acquired training dataset. Further information can be gathered from the BraTS challenge¹. The rest of the paper is organized as follows, section II describes the literature survey of deep learning related research, section III describes the proposed cascaded GAN based solution with discriminator loss function, section IV deals with Experimental section and results, followed by section V describing conclusion.

II. LITERATURE REVIEW

In this section we study some of the recently proposed methods which have significantly accelerated the research in brain tumor segmentation. CNNs have outperformed the existing image processing based conventional methods by their ability to effectively learn hierarchy of features at multiple scales from data. A fully convolutional network (FCN) for semantic segmentation was first proposed by Long et al. [5]. In this work, a fully connected layer which is generally found at the end of network was replaced by convolutional layer to obtain a coarse label map.

A U-net like structure was presented by Ronneberger et al. in [9] for the purpose of segmenting neuronal structures in electron microscopic stacks. With idea of residual connections, originally inspired from [4], U-net made a drastic jump in the performance. In addition to this, Havaei et al. [3] introduced the idea of cascaded CNNs. The main idea behind this architecture is to use output probabilities of previous stage as additional input for the next stage. Instead of full image, it takes patches of image as input. From the detailed study of BraTS 2017 proceedings², we developed various new insights of looking at the segmentation problem. Based on our study, some of the important factors to be considered in order to achieve good performance are discussed here. Instead of using a single path in the CNN, using multiple pathways of different resolutions helps to capture the both short and long distance

¹BraTS Data : <https://www.med.upenn.edu/sbia/brats2017/data.html>

²BraTS Proceedings: https://www.cbica.upenn.edu/sbia/Spyridon.Bakas/MICCAI_BraTS/MICCAI_BraTS_2017_proceedings_shortPapers.pdf

pixel correlations in a better way. Another important factor is the 3-dimensional shape of brain tumor. So taking 3D patches as input captures 3D-dimensional pixel correlations compared to only 2-dimensional correlation when 2D slices are taken as input. The choice of loss function has a significant impact on the performance because it eventually decides how the network learns during training. We have employed a multi-modal loss function in our proposed method inspired from GAN which will be discussed in detail in later sections.

III. METHODOLOGY

The Magnetic Resonance Images (MRI) that we obtain are outputs of Radio Frequency signals processed over patients at the diagnostic center. These MR images have various irregularities in them viz. bias or intensity inhomogeneities. The reason being the imperfect radiation patterns of the perceiving (receiving) antennas and their location in the Magnetic Resonance images. The result of such distortions affects basically the intensity values lying on edges as when compared to the centre. We employ two different prepossessing steps to overcome the above mentioned limitations, followed by z-score normalization.

A. Pre-processing

1) *N4 bias correction*: It [10] maximizes the high frequency content of the intensity distribution by smoothing the multiplicative field. It considers an ideal image u corrupted with bias field f and noise n , thus giving an image v as below,

$$v(x) = u(x)f(x) + n(x) \quad (1)$$

Neglecting noise term and taking logarithm on both sides,

$$\hat{v}(x) = \hat{u}(x) + \hat{f}(x) \quad (2)$$

An iterative solution can be proposed now for deriving the uncorrected image, as given below,

$$\begin{aligned} \hat{u} &= \hat{u}^{n-1} - \hat{f}_r^n \\ &= \hat{u}^{n-1} - S^* \{ \hat{u}^{n-1} - E[\hat{u}|\hat{u}^{n-1}] \} \end{aligned} \quad (3)$$

where $S^* \{ \cdot \}$ is a Spline approximator, \hat{f}_r^n is the estimation of the bias field in n^{th} iteration and $E[\hat{u}|\hat{u}^{n-1}]$ is the actual image expected value given the current estimation.

2) *Intensity Correction*: For each of modalities present in BraTS dataset we perform an intensity correction due to non-uniform intensity ranges covering across all the patients in a particular modality. This intensity standardization method was proposed by Nyul et al. [8]. This method picks from each sequence a set of intensity landmarks, $\text{Intensity}_{\text{landmarks}} = \{ pc_1, i_{p10}, \dots, i_{p90}, pc_2 \}$ values were chosen according to [7]. After training gets accomplished, we map the original intensities between two landmarks into corresponding learned landmarks. In this way we make the histogram of each modality to be similar among the patients i.e. histogram equalization is also achieved here.

B. Proposed Method : using GANs

Generative Adversarial Networks first proposed in [2] are the state-of-the-art architecture in computer vision and NLP related research topics introduced by Goodfellow Ian et al. Though a GAN based solution SeGAN recently published by Yuan et al. in [12], has achieved the state-of-the-art result but it takes only the discriminator to generate the loss function. It neglects a key property of discriminating generator based solution with ground truth by taking adversarial loss. To design the proposed GAN based model we take inspiration from the positive output that we received from the two papers presented namely Anisotropic CNN by Wang et al. in [11] and Patch based UNet by Beers et al. in [1], [11] being the runner-up in BraTS 2017 Challenge and Patch based techniques have previously won the competition. Since U-Net based architecture are widely used for providing segmented output of an entire 240×240 image in one forward pass we build our generator based on U-Net like model. Also as Anisotropic CNN does execution of Whole Tumor, Tumor core and Enhancing Tumor in a sequential order we deploy this idea on our GAN based model.

1) *Cascaded Structure*: We employ three (Generator-Discriminator) structures in a cascaded form to classify the three labels i.e. Whole Tumor (WT), Tumor Core (TC) and Enhancing Tumor (ET). The network is shown in Fig. 1 where the input to the first network i.e. WTGAN is of shape $240 \times 240 \times 4$, where 240×240 is the image slice dimension and 4 being the number of modalities. The output of the WTGAN is of shape 240×240 which is the probability map of each pixel being classified as Whole Tumor or not. This whole tumor region is then cropped to be fed as an input to the TCGAN (Tumor Core GAN). Evidently, the input shape of TCGAN is a variable depending on the output from WTGAN and it is also a binary classifier output of whose depicts either a Tumor Core (TC) or a Whole Tumor (WT). The output region of this TCGAN is then cropped for TC region only and is fed to the input of Enhancing Tumor (ET) GAN. Thus forming a cascaded structure of three separate GANs namely WTGAN, TCGAN, ETGAN, as shown in Fig. 1. The Generator and discriminator network used for all the three GAN variants are kept same and are explained in the subsequent sections.

2) *Generator Network*: The Generator network is an U-Net based model shown in Fig 2. The network comprises of downsampling layers (encoding layers) followed by upsampling layers (decoding layers). The (Encoding-Decoding) structures forms a U like shape thus justifying the name U-Net. The U-Net network comprises of multiple (Convolution-Batch Normalization-Pool) layers in the encoder structure while (Up-Pool-Concat-Convolution-Batch Norm) layers in the decoding structure. The concatenation layer is for adding the residual skip connections forming a linkage between the coding and decoding structures. The skip connections are used for the information that was captured in the initial layers and which are required for reconstruction during the up-sampling. So the information that we had in the primary layers can be

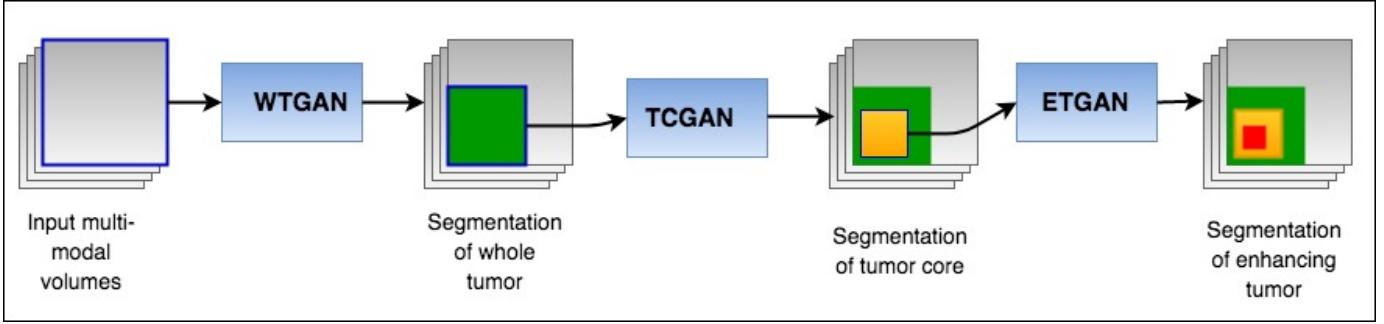


Fig. 1. Cascaded Generative Adversarial Network Structure

fed explicitly to the later layers using the skip architecture. The convolutional layer consists of two 3×3 kernel filter size followed by ReLU activation layer, batch normalization layer and then max pooling with kernel size 3×3 of stride 2×2 . This block is repeated three times in encoding layer. With similar filter sizes the decoding layer performs operation in reverse along with skip connections. Finally the output of the Generator is a binary classifier of shape same as that of input but output depth being only one for binary classification.

3) Discriminator Network: The input to the i^{th} ($i = 1$ represents WTGAN and so on) discriminator is of shape $W_i \times H_i \times c$ where W_i and H_i are the output shapes of i^{th} Generator and $c = 4$ represents the number of modalities. The Generator output (i.e. probability map) and ground truth are multiplied with multi-modal input image to form the input for discriminator corresponding to fake class and real class respectively. The Discriminator network is then formed as shown in Fig. 2. The network consists of repeated (Conv-ReLu-BN) layers with a repetition factor of 3 followed by two Dense layer and finally a sigmoid layer to predict whether the given output is from a real(input from ground-truth) or fake class (input from Generator).

4) Loss Function: In spite of relying only on the final sigmoid layer cross-entropy loss function we propose a novel loss function capable of capturing adversarial losses provided by the discriminator and content loss that we derive from the j^{th} output layer of discriminator where j corresponds to 1, 2 and 3 (the repetition factor of D being 3). After the Generator outputs a predicted image $I_{\text{predicted}}$, it is multiplied with 4 input modality images to give a masked I_{PM} image corresponding to predicted Generator output. Thus each I_{PM} of initial shape $W_i \times H_i$ now becomes $W_i \times H_i \times n$ as shown in 4,

$$I_{\text{PM}} = I_{\text{predicted}} \times I_{\text{modality}}. \quad (4)$$

A similar operation is done for ground truth label being multiplied with the 4 modality images to give I_{GT} as output as follows,

$$I_{\text{GTM}} = I_{\text{GT}} \times I_{\text{modality}} \quad (5)$$

Thus the optimization function of generator can be written as,

$$\hat{\theta}_G = \operatorname{argmin}_{\Theta_G} \frac{1}{N} \sum_{n=1}^{n=N} \text{Loss}(I_{\text{PM}}, I_{\text{GTM}}). \quad (6)$$

$$\text{Loss}(I_{\text{PM}}, I_{\text{GTM}}) = \alpha \times \text{Content} + \beta \times \text{Adversarial}. \quad (7)$$

The parameters α and β are the weights we choose to provide to the content and adversarial loss respectively. Since this is a pixel level segmentation task hence we give more weight to the content as compared to adversarial. The weights we choose are 100:1. The content and adversarial loss now be given as follows.

a) Adversarial Loss: The adversarial loss is defined based on Discriminator's probability output so that the discriminator pushes the generator to produce image which bear the probabilistic distribution of more natural looking segmented tumor output.

$$\text{AdversarialLoss} = \sum_{n=1}^{n=N} [\log(1 - D_{\theta_D}(G_{\theta_G}(I_{\text{modality}}))] \quad (8)$$

Here $D_{\theta_D}(G_{\theta_G}(I_{\text{modality}}))$ is the probability that the reconstructed image $G_{\theta_G}(I_{\text{modality}})$ is actually a correct representation of the ground truth label.

b) Content Loss: The content loss as defined in Equation 3.9 is taken as the pixel wise difference between the three CNN output layers from the discriminator network. In our case, since the content loss accounts for multimodality dependencies between feature maps, we also call it Multi-scale feature loss shown in Fig 3.

$$\text{ContentLoss} = \frac{1}{W_i \times H_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (\phi_i(I_{\text{GTM}})_{x,y} - \phi_i(G_{\theta_G}(I_{\text{PM}}))_{x,y})^2 \quad (9)$$

Here, W_i and H_i represents the width and height of i^{th} Convolution layer output, $\phi_i(\cdot)$. The pixel wise difference is then squared to take mean square loss over the entire output image from each layer. We then define for the discriminator network D_{θ_D} which is optimized along with Generator network G_{θ_G} to solve the min-max problem given below, where $E[\cdot]$ is the expectation over the training samples,

$$\begin{aligned} & \min_{\theta_G} \max_{\theta_D} E[\log D_{\theta_D}(I_{\text{GTM}})] \\ & + E[\log(1 - D_{\theta_D}(G_{\theta_G}(I_{\text{PM}})))] \end{aligned} \quad (10)$$

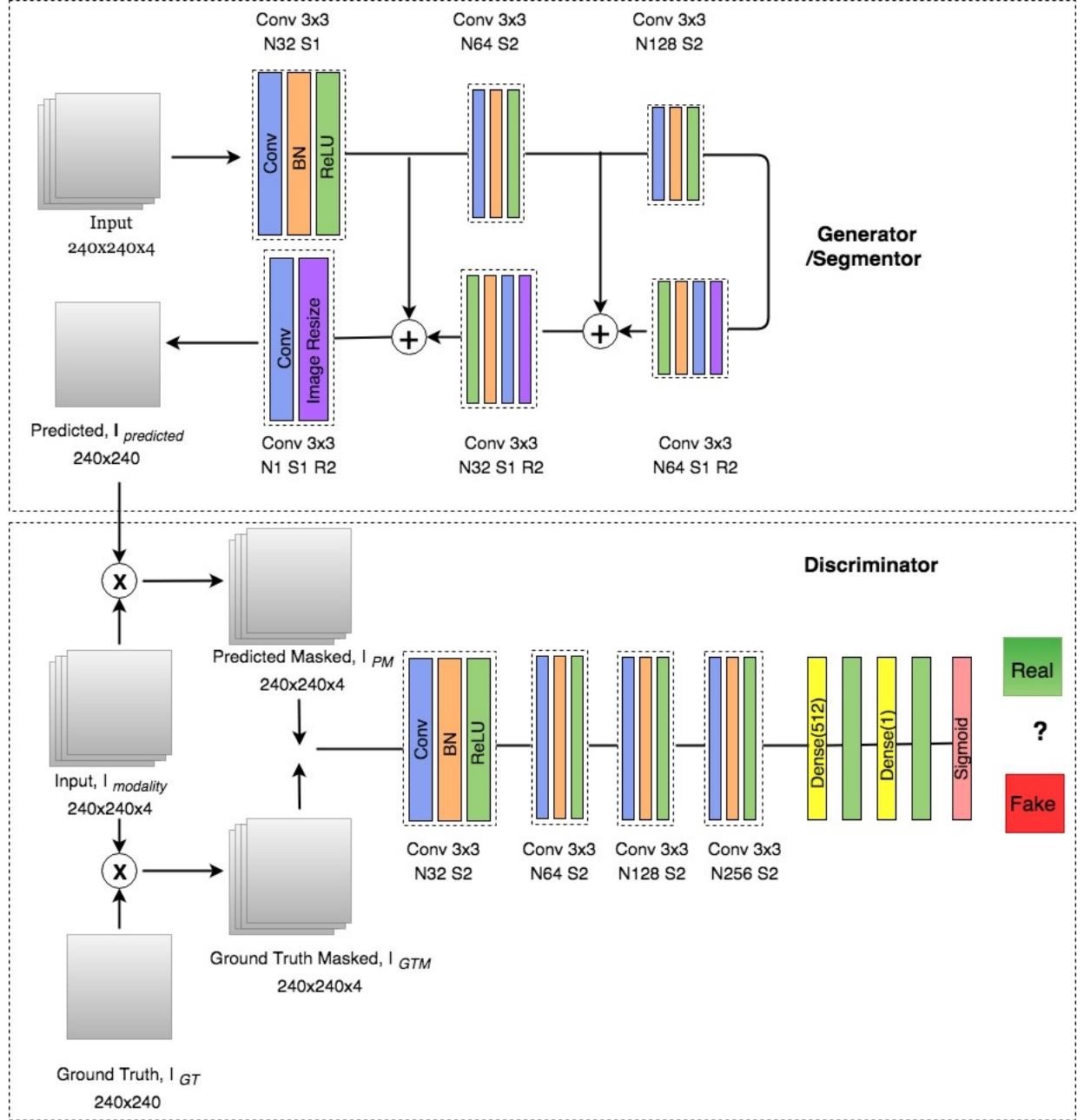


Fig. 2. Proposed architecture of Generator and Discriminator Network

IV. EXPERIMENT AND RESULTS

The dataset [6] that we used for all our experimental purpose is provided by MICCAI BraTS Challenge for the year 2017. This dataset consists of MRI scans of 210 patients suffering from High Grade Gliomas (HGG). Each patient has 155 slices forming up the 3D space of a human brain. The size of each slice 240×240 forms a 2D image. Also for each patient a total of 4 different modality slices are provided as already mentioned in introduction along with the ground truth labelling of the tumor regions. For having a uniform comparison among all the methods the test-train split is also

taken alike. For training purpose we randomly selected 150 patients out of 210 thus forming our train dataset while the test data consists of remaining 60 patients on which testing have been performed. During training the neural networks, 80 : 20 validation split has also been performed on the training dataset without indulging the testing data. To compare the proposed methods results with the Anisotropic [11] and Patch based models [1] for visualization purpose we select a particular patient namely Brats17_CBICA_AUR_1 (slice 80) considered for showing results of all the three methods. All the models are implemented in Keras with Tensorflow backend and NiftyNet, a medical imaging library based on keras and tensorflow. The

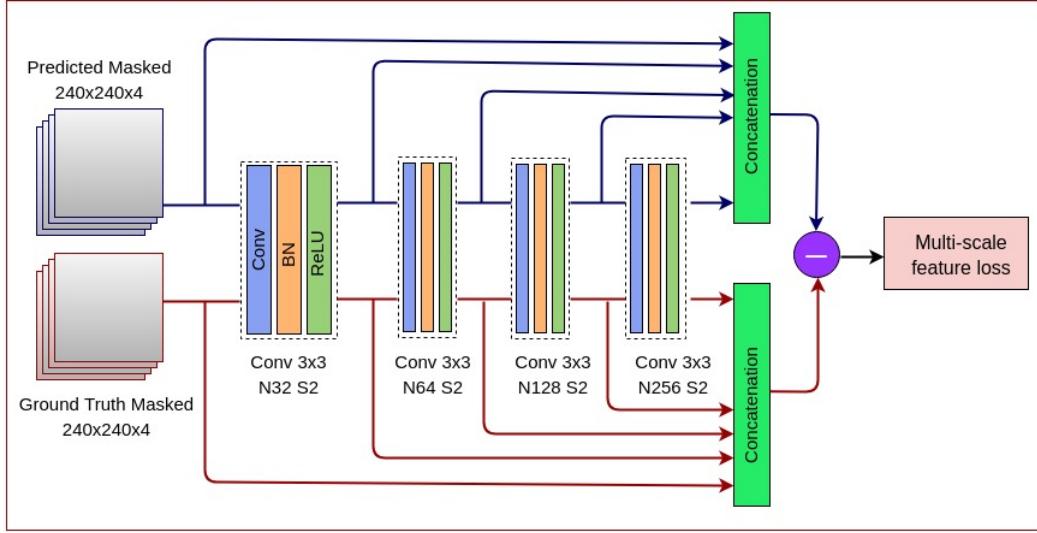


Fig. 3. Discriminator Network showing Multi-scale Multi feature loss

experiments are performed on a Nvidia Titan X GPU with 6GB memory and CPU based internal RAM with 8GB memory.

The input shape in our case is taken as $240 \times 240 \times 4$ for WTGAN and subsequent cropping(rectangular boundaries) of the predicted image (based on WTGAN output) as well as cropping on the original ground-truth image with WT region. This is fed to TCGAN and likewise for ETGAN. The Optimizer used in our model is the Adam Optimizer [13]. Dropout [9] is introduced in Generator Model after every (Conv-BN-ReLU) layer during the encoder block. The dropout probability value is chosen to be 0.5. Two dense layers of activation neurons of size 512 and 1 are chosen respectively in the Discriminator network, which are in fact the only fully connected layer in the entire model. The loss function describes previously uses a Content loss (MSE based loss) which is summed over a batch size of 4. The number of training epoch was restricted to 30.

A. Comparison of Segmented results

In this section we compare the results of Anisotropic and Patch based models against our proposed architecture both visually and statistically to get a better view of all the three methods and their segmentation results. The contrasting results from a randomly selected patient can be observed in Fig. 4 where Anisotropic Cascaded CNN outperforms the patch based network due to the sequential execution of three tumor regions thus reducing false positives. In the patch based model the cascaded structure of (whole Tumor, Tumor core, Enhancing Tumor) is nowhere taken into consideration. Comparing our GAN based approach to the anisotropic method we observe that the anisotropic method due to its singular cross entropy loss function present at the output of the network, the network tries to smoothen the classification output, but considering our multi-scale loss i.e. summation of the difference between various network output layers and the discriminator forcing our generator to reach close to the original distribution

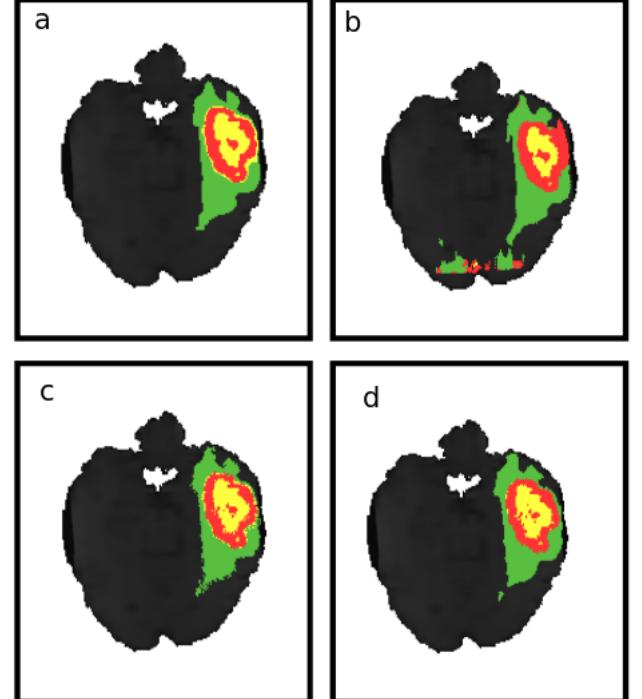


Fig. 4. Comparison of segmentation results for patient Brats17_CBICA_AUR_1. (a) Anisotropic CNN (b) Patch-based3D U-Net (c) Cascaded GAN proposed (d) Ground truth

of dataset labelling, so that it resembles the most with the ground truth.

B. Statistical Comparison of the models

The BraTS Challenge provides a standard method of comparison by taking Dice score, PPV score and Specificity score for all the three regions viz. WT, TC and ET. The three metrics are given as follows,

$$Sensitivity = \frac{TP}{TP + FN} \quad Specificity = \frac{TN}{TN + FP}$$

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN}$$

where TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative. The table below shows the mean scores for all the performance metrics for every method. From the three tables shown below we note that Dice score and Sensitivity values are near to each other asserting that the FP rate is very low, which should follow in our case as FP means asserting a healthy tissue to be of tumor region. Likewise the very high absolute value of Specificity tells us that the segmentation method gives high TN rate i.e. if it is a brain tissue than there is a very high rate of segmenting it as a brain tissue only. Further we observe that in all the three tables our method perform better in comparison to patch based solution for ET and TC region. The reason is because of taking cascaded structures in consideration. For the WT region Anisotropic CNN performs better as it is trained on a 3D architecture and Whole Tumor being the first layer, the cascaded architecture design cannot be taken into consideration here. The below tables highlights in bold the peak value of Dice (Table 1), Sensitivity (Table 2), Specificity (Table 3) scores for the best model.

TABLE I
DICE SCORES

Methods	ET	WT	TC
Anisotropic CNN	0.783	0.874	0.775
Patch Based U-Net	0.751	0.856	0.739
Cascaded GAN(proposed)	0.820	0.869	0.783

TABLE II
SENSITIVITY

Methods	ET	WT	TC
Anisotropic CNN	0.775	0.912	0.841
Patch Based U-Net	0.735	0.892	0.781
Cascaded GAN(proposed)	0.791	0.905	0.804

TABLE III
SPECIFICITY

Methods	ET	WT	TC
Anisotropic CNN	0.9949	0.9942	0.9974
Patch Based U-Net	0.9765	0.9843	0.9903
Cascaded GAN(proposed)	0.9985	0.9879	0.9964

V. CONCLUSION

The proposed Cascaded Generative Adversarial Network-based solution has U-Net like generator to reduce the training time and trainable parameters as it uses Fully Convolutional Network instead of Fully Connected layers. We train in a

cascaded fashion viz. firstly Whole Tumor, then Tumor Core and finally Enhancing Tumor which helps in reducing the false positives. It is because of this cascaded training we can see a major improvement in the results for Enhancing and Tumor core against other methods as these two networks have inputs delivered from Whole Tumor Network (the first network). Also we shift from a categorical cross entropy loss function to a more robust Discriminator based loss where two losses are taken, firstly the adversarial loss to enhance the Generator produced output as close to the ground truth image as possible and secondly the content loss which takes pixel-wise difference among the intermediate layers of Discriminator, this helps in the reduction of overly smooth contour boundaries. Finally, the learned network shows the improved values of Dice, PPV and Specificity scores for Enhancing and Tumor Core over the existing methods.

REFERENCES

- [1] Andrew Beers, Ken Chang, James Brown, Emmett Sartor, CP Mammen, Elizabeth Gerstner, Bruce Rosen, and Jayashree Kalpathy-Cramer. Sequential 3d u-nets for biologically-informed brain tumor segmentation. *arXiv preprint arXiv:1709.02967*, 2017.
- [2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [3] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical image analysis*, 35:18–31, 2017.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [6] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993, 2015.
- [7] László G Nyúl and Jayaram K Udupa. On standardizing the mr image intensity scale. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(6):1072–1081, 1999.
- [8] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. New variants of a method of mri scale standardization. *IEEE transactions on medical imaging*, 19(2):143–150, 2000.
- [9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [10] N Tustison and J Gee. N4itk: Nicks n3 itk implementation for mri bias field correction. *Insight Journal*, 9, 2009.
- [11] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI Brainlesion Workshop*, pages 178–190. Springer, 2017.
- [12] Yuan Xue, Tao Xu, Han Zhang, L Rodney Long, and Xiaolei Huang. Segan: Adversarial network with multi-scale l1 loss for medical image segmentation. *Neuroinformatics*, pages 1–10, 2018.

Decision Support System for Liver Cancer Diagnosis using Focus Features in NSCT Domain

Lakshmipriya Balagourouchetty
Department of Electronics and Communication Engineering
Pondicherry Engineering College
Puducherry, India
lakshmipriya@pec.edu

Jayanthi.K.Pragatheeswaran
Department of Electronics and Communication Engineering
Pondicherry Engineering College
Puducherry, India
jayanthi@pec.edu

Biju Pottakkat
Department of Surgical Gastroenterogy
JIPMER
Puducherry, India
bijupottakkat@gmail.com

Rammkumar Govindarajalou
Department of Radio Diagnosis
JIPMER
Puducherry, India
Gramk80@gmail.com

Abstract— Diagnosis of liver cancer by medical experts using imaging modalities is found to be sub-optimal as different lesions exhibit similar visual appearance in the spatial domain. Thus computer aided diagnostic tools play a significant role in providing a decision support system for radiologists to minimize the risk of false diagnosis. This paper proposes a different feature set using focus operators for classifying different classes of liver cancer. As computation of focus measure involves the local neighborhood of pixel, focus operator is believed to indirectly measure the intricate texture details of the image. This knowledge of focus operator is exploited in NSCT domain to capture the directional components as feature variables replacing the classic texture features. The results in terms of classification accuracy and kappa coefficient proclaim that the focus operators can be employed as feature variables for classification scenario as it outperforms the state-of-the art texture features.

Keywords—focus measure, feature extraction, feature selection, outliers, classification.

I. INTRODUCTION

Multiphase contrast enhanced Computed Tomography (CT) is the widely preferred imaging modality by the physicians for the detection and diagnosis of different types of liver tumour namely hepatocellular carcinoma (HCC), the primary liver cancer; hemangioma (HEM) and so on. Multiple phases of liver CT implies CT acquisition at four different time intervals: i) Plain or unenhanced CT image (without contrast injection) ii) arterial phase CT (after 30 – 40 seconds of contrast injection) iii) portal venous phase CT (after 60-70 seconds of contrast injection) and iv) washout phase CT (2 – 3 minutes of contrast injection) [1]. Liver tumour tissues exhibit different enhancement in different phases of CT images and this visual enhancement information are collectively utilized by the medical experts for the diagnosis of liver lesions. Although high quality CT images are available in recent years, in some cases where the enhancement pattern is obscure, clinicians depend upon the tissue biopsy to confirm their diagnosis. Thus there is

a growing need for the automated diagnostic tool to figure out the liver lesion with appropriate interpretation. Computer aided diagnosis (CAD) of liver related pathologies provides a decision support system for both radiologists and medical experts by eliminating the need for tissue biopsies for confirming their diagnosis.

A variety of CAD systems is available in literature for the classification of liver lesions and most of them are based on texture and shape features. Texture features namely Gray Level Co-occurrence Matrix (GLCM) features are predominantly used in literature for the analysis and classification of various types of tumours [1 - 6]. In addition to GLCM features, other texture features like Gray Level Run Length Matrix (GLRLM) features, fractal features, Laws texture energy measures, Gabor features, gray level difference statistics (GLDS) features, neighborhood Gray Tone Difference Matrix (NGTDM) features are also adopted for the classification of pathology from images of different modalities like CT, ultrasound (US) and Magnetic Resonance Imaging (MRI). Gleatos *et.al* [2] have used GLCM features coupled with dimensionality reduction using Sequential Forward Selection (SFS) and Genetic Algorithm (GA) and achieved 97% classification accuracy using Neural Network (NN) classifier. In [3], statistical texture (GLCM) features in wavelet domain along with feature selection using Sequential Forward Floating Search (SFFS) and GA were employed for the classification of cirrhosis and fatty liver classification using NN and achieved higher accuracy when compared to techniques in spatial domain. In [4], GLCM features coupled with kinetic shape features have been used for the categorization of liver lesions using logistic regression classifier. Classification of benign liver tumours, cyst and HEM was carried out in [5] using statistical features extracted from the portal venous phase images using Support Vector Machine Classifier (SVM). Although texture analysis in spatial domain have reported an impressive performance in lesion classification of various tumours, researchers have also extracted features in a transform domain and achieved higher classification accuracy than the customary spatial domain techniques. Biorthogonal wavelet based statistical texture feature extraction paired with GA based feature selection [6] has reported better accuracy in classifying cirrhosis and fatty

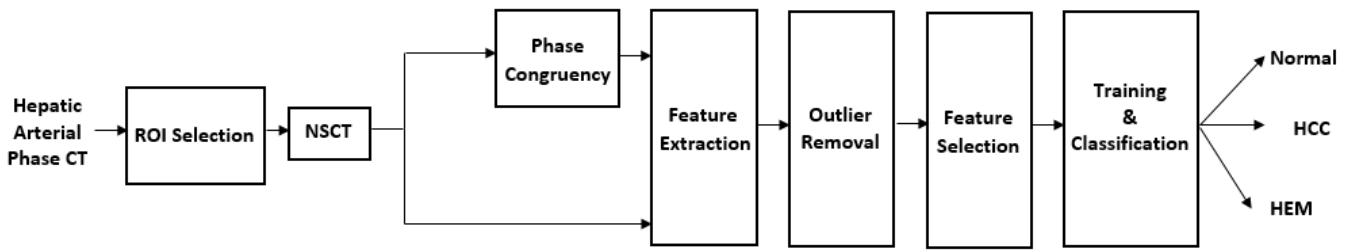


Fig.1 Proposed CAD System for Liver Tumour Diagnosis

liver when compared to the same scheme in conventional gray level domain. Authors of [7] have attained phenomenal results with the help of entropy features extracted from the Curvelet transform coefficients of the fatty liver ultrasound images followed by Locality Sensitive Discriminant Analysis (LSDA) feature reduction method. Wavelet based shape and texture feature extraction for the classification of dermoscopy images was proposed in [8] and achieved an accuracy of 92.61% and specificity of 91%. An hybrid approach proposed in [9] for the classification of focal and diffused liver lesions using biorthogonal wavelet transform based GLRLM texture features in conjunction with random forest classifier has achieved superior performance over the spatial domain counterpart. Contourlet and Non-Subsampled Contourlet Transform (NSCT) based texture analysis for the breast tumour classification was carried out in [10 – 12] and [13 – 15] respectively. It is evident from the above review that an exhaustive research has been carried out using texture, statistical features for the classification of medical images both in spatial and in different multiscale geometric domains. Hence, in this paper a novel investigation approach using focus measures based feature extraction using NSCT is carried out to provide an alternate option for the lesions classification. Generally, focus measure denotes the measure of image clarity. The focus measures which are used to evaluate the sharpness of a pixel by considering the local neighborhood of the center pixel are used as features in the proposed CAD system. In addition to this, feature optimization is also performed using max – relevance and min – redundancy criterion approach. These are the major highlights of this work.

II. MATERIALS AND METHODS

A. Data Acquisition

The images used for this study were obtained from Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), Puducherry, India, after the approval of JIPMER Scientific Advisory Committee (JSAC) and Institute Ethics Committee (IE). CT scans were acquired using Siemens Sensation 64 detector scanner via a standard four phase contrast enhanced imaging protocol. The database used in this research work comprises of 400 CT images as presented in table II. The

hepatic arterial phase images which are predominantly referred by radiologists for diagnosis were used in the study. The pathology types of the lesions were ratified through clinical features and visual examination of all 4 phases of CT images by an experienced gastroenterologist and a radiologist.

B. Methodology

In this paper, a CAD system to categorize the input liver lesion sample under three categories, namely: Normal or Abnormal and if it is abnormal it is further probed to classify either as benign (HEM) or malignant (HCC), the block diagram of which is presented in figure 1.

The lesions extracted manually from the hepatic arterial phase images corresponding to normal liver, HCC and HEM were fed as input to the CAD system. The input samples were decomposed using NSCT which is a multi-resolution image decomposition framework to decompose the source image into one low frequency sub-band coefficient and a series of high frequency sub-band coefficients. The number of high frequency sub-images generated depends upon the decomposition levels. The low frequency coefficients of the NSCT decomposition normally called as approximation coefficients correspond to the contrast contents of the source image sample and the high frequency coefficients reflect the edge information and the texture pattern of the source image. In this work, number of decomposition levels in NSCT is chosen to be 1. The contrast contents in the low frequency sub-band coefficients are dependent on the imaging modality, the amount of contrast injected into the patient and the machine settings at the time of CT acquisition. Hence the contrast of the liver parenchyma tends to exhibit different pixel intensity mapping for different patients though the imaging modality is same.

To avoid this problem of having different contrast for the similar tissues, contrast content which is independent of pixel mapping is preferred and this is achieved by computing two dimensional phase congruency (PC) of the low frequency sub-band coefficients. Phase congruency is a contrast invariant feature formulated in [16] in continuance with the result produced by Oppenheim and Lim [17] that the phase information of the image carry significant details than the intensity level or amplitude of an image. Following the feature extraction, 7 best features were selected using max – relevance and min – redundancy criterion based Pearson's correlation coefficient (RRPC) method of feature selection [18]. Finally

training and classification of hepatic lesions were done using SVM classifier.

C. Feature Extraction

Focus measure operators were used in the literature as a measure of image clarity to evaluate the degree of focus. A variety of focus measures are available in literature and these measures are utilized in evaluating the performance of certain image processing schemes other than its customary applications. The focus measure operators can be categorized into six different types [19]:

1. Gradient based operators, wherein the operators are computed as the first derivative of the image.
2. Laplacian based operators: Computed as the second derivative of the image to automatically detect the edges of the image.
3. Wavelet based operators: These operators are computed by means of Discrete Wavelet Transform (DWT) to analyze the spatial and frequency components of the image.
4. Statistics based operators: These operators make use of statistical property of the image for its computation.
5. DCT based operators: Analogous to wavelet based operators, these operators make use of Discrete Cosine Transform (DCT) coefficients for the computation of focus depth of an image.
6. Miscellaneous operators: encompasses all the focus operators which do not fall under any of the above five categories.

TABLE I. FOCUS FEATURES

Sl No	Focus Feature	Abbreviation
1	Absolute Central Moment	ACM
2	Brenner's Focus Measure	BREN
3	Image curvature	CURV
4	Energy of Gradient	EoG
5	Energy of Laplacian	EoL
6	Spatial Frequency	SF
7	Sum Modified Laplacian	SML
8	Diagonal Laplacian	DL
9	Variance of Laplacian	VoL
10	Tenengrad	TG
11	Tenengrad Variance	TGV

The fact that the focus measure operators exhibit difference in intensity levels and are computed by examining the irregularities in image texture due to surface orientation [20], makes these operators suitable to characterize the inherent properties of image texture. This is being utilized in this work to perform classification based on focus features which has not been reported in literature to the best of the knowledge of the authors. A total of 11 such focus measures as listed in table I along with the corresponding abbreviations were computed for all the high and low frequency sub-band coefficients of the NSCT transformed image sample.

ACM [21] is a histogram based focus measure extensively used to quantify the quality of the image.

$$ACM = \sum_{k=1}^L |k - \mu| P_k \quad (1)$$

μ is the mean intensity, L is the number of graylevels and P_k is the relative frequency of kth graylevel. The feature BREN is

computed based on the second difference of the gray levels of an image. Mathematically, it is given by

$$BREN = \sum_x \sum_y |I(x, y) - I(x + 2, y)|^2 \quad (2)$$

CURV [22] is another focus measure computed by considering the gray values as 3D surface ie.(x,y,g(x,y)) where g(x,y) represents the pixel intensity at the location (x,y). As a first step, approximation of the surface is done as $f(x) = c_0x + c_1y + c_2x^2 + c_3y^2$, where the coefficients $C = [c_0 \ c_1 \ c_2 \ c_3]^t$ are found using least square approximation with M_1 and M_2 as given (4).

$$C = \left[M_1 \cdot I, \ M_1^T \cdot I, \ \frac{3}{2} M_2 \cdot I - M_2^T \cdot I, \ \frac{3}{10} M_2 \cdot I - \frac{M_2 \cdot I}{5} \right]^t \quad (3)$$

where

$$M_1 = \frac{1}{6} \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \text{ and } M_2 = \frac{1}{5} \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \quad (4)$$

Subsequently, these coefficients are simply combined to obtain the image curvature focus measure as
 $CURV = |c_0| + |c_1| + |c_2| + |c_3| \quad (5)$

EoG [23] is computed by calculating the squared values of image gradient as

$$EoG = \sum_x \sum_y (I_x^2 + I_y^2) \quad (6)$$

Where

$$I_x = I(x + 1, y) - I(x, y) \text{ and } I_y = I(x, y + 1) - I(x, y) \quad (7)$$

EoG is a good focus operator widely used to reduce the noise and artifacts in deburred images and is also used as a metric to evaluate the image focus quality. EoL [23] is an additional focus measure to examine the impact of high spatial frequency linked with image border sharpness.

$$EoL = \sum_x \sum_y \Delta I(x, y)^2 \quad (8)$$

where ΔI is the image Laplacian obtained by convolving I with the Laplacian mask. Spatial frequency is an improvised version of EoG used in the selection of high frequency coefficients in image fusion applications and also in quantifying the edge strength of the image. The mathematical computation of this parameter is given as

$$SF = \sqrt{RF^2 + CF^2} \quad (9)$$

where RF and CF are row and column frequency respectively.

$$RF = \sqrt{\frac{1}{MN} \sum_{x=1}^M \sum_{y=2}^N [I(x, y) - I(x, y - 1)]^2} \quad (10)$$

$$CF = \sqrt{\frac{1}{MN} \sum_{x=2}^M \sum_{y=1}^N [I(x, y) - I(x - 1, y)]^2} \quad (11)$$

SML operator is basically a focus measure to pixel variation in an image. It measures the sharpness in the contrast contents of the image and is calculated as

$$SML = \sum_{i=x-N}^{x+N} \sum_{j=y-N}^{y+N} \nabla_{ML}^2 f(i,j) \text{ for } f(i,j) \geq T \quad (12)$$

Where T is a discrimination threshold value.

$$\nabla_{ML}^2 f(x,y) = |2f(x,y) - f(x-1,y) - f(x+1,y)| + |2f(x,y) - f(x,y-1) - f(x,y+1)| \quad (13)$$

VoL [24] expressed in (14) measures the local variance of gray levels in autofocus applications.

$$VoL = \sum (\Delta I(i,j) - \bar{\Delta I})^2 \quad (14)$$

Where $\bar{\Delta I}$ is the mean value of the image Laplacian. DL [25] is a mathematical operator formulated by considering the diagonal neighborhood in addition to the horizontal and vertical ones in the computation of SML. TG and TGV were computed by Tanenbaum by obtaining the gradient magnitude of Sobel operator.

$$TG = \sum_{i \in x} \sum_{j \in y} G_x(i,j)^2 + G_y(i,j)^2 \quad (15)$$

G_x and G_y are image gradients along rows and column respectively.

$$TGV = \sum_{i \in x} \sum_{j \in y} (G(i,j) - \bar{G})^2 \quad (16)$$

where \bar{G} is the mean value of the gradient magnitude.

The separation of the contrast and edge contents of input image sample being the main objective of transformation using NSCT, decomposition level for NSCT is limited to one in this work. Thus, one LF sub-band coefficients and two HF sub-band coefficients will be available at the output of single stage NSCT decomposition giving rise to a sum of 33 focus features extracted from every sample of hepatic CT image. The computation of these focus parameters are detailed in [20-25].

D. Outlier Removal

Following, the feature extraction stage, the outlier values are removed from extracted feature data set using Whisker box plot method. Outliers are the feature variable values with a value lying below or above with 1.5 times the inter quartile range. Whisker's box plot displays the graphical statistical summary of a variable in the form of plot. The cases corresponding to the outliers are removed from the feature set. Figure 2 shows the sample whisker's box plot with ACM, SML and DL features extracted from the LF sub-band of feature set of normal class. No outlier cases are found for ACM and outliers are present above the box in SML and DL features and they are excluded from feature set.

E. Feature Selection Using RRPC Coefficient

After the removal of outliers, relevant features out of the extracted 33 features were selected using RRPC coefficient. RRPC is a forward filter based feature reduction technique with reduced computational complexity compared to the

conventional filter methods. This method is essentially an incremental search technique based feature selection to achieve better trade-off between relevant and redundant features. Pearson's correlation coefficient is used as a means to measure feature-to-feature information and feature-to-label information. The effectiveness of RRPC algorithm is demonstrated in [18] and is as such utilized in this study for the purpose of selecting highly relevant feature subset to achieve better classification accuracy.

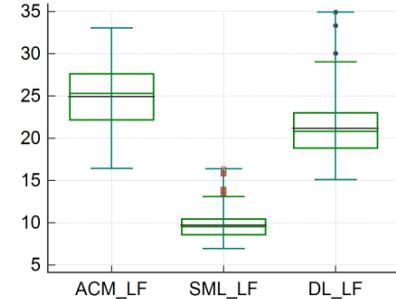


Fig.2 A sample boxplot demonstrating the outlier values in ACM, SML and DL features extracted from the LF sub-bands

F. SVM Classifier

Classifier is the last module of the CAD system which is anticipated to categorize the type of hepatic lesions as normal, benign (HEM) or malignant (HCC); and a multiclass linear SVM is used for this purpose in this work. SVM classifier is originally developed based on the principle of minimizing the training set error and maximizing the functional margin between the two classes under consideration. Since SVM has reported an exceptionally better performance in classification process, this classifier was used to evaluate the performance of the focus measure features chosen in this study.

III. RESULTS AND DISCUSSION

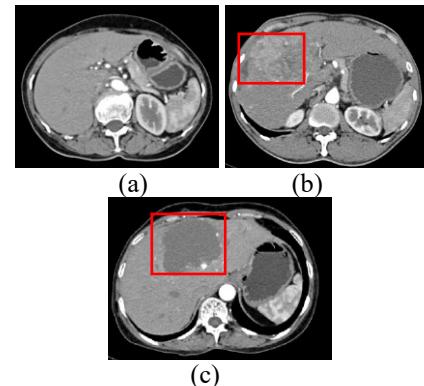


Fig. 3 Liver CT Images (a) Normal Liver (b) HCC Liver (c) HEM liver

Liver CT images corresponding to normal liver, HCC and HEM affected liver in arterial phase are shown in fig.3 with the unhealthy portion of liver highlighted using a red colored box. Prior to classification, the regions of interest (ROIs) corresponding to normal region and lesions are extracted

manually from the CT images by two of the authors, an experienced surgeon and an experienced radiologist. The sample ROIs corresponding to the three classes of hepatic images namely, healthy liver, HCC liver and HEM liver are presented in fig.4. The seven best features selected using RRPC feature selection are tabulated in table II in the order of highest relevance.

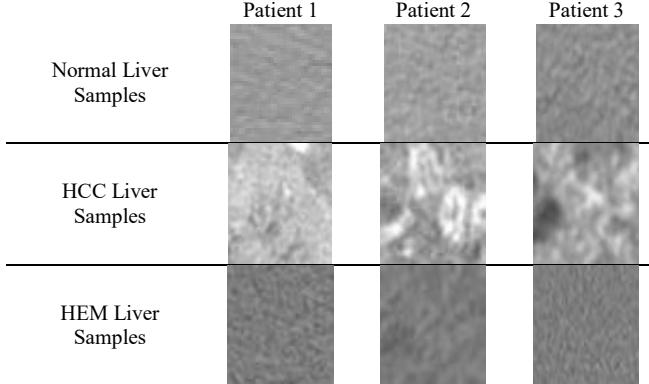


Fig.4 Sample ROIs of hepatic tissues

TABLE II. LIST OF BEST 7 FEATURES SELECTED USING RRPC ALGORITHM

SI No	Feature	Abbreviation	Sub-band
1	Absolute Central Moment	ACM	HF-3
2	Tenengrad variance	TV	HF-1
3	Energy of Laplacian	EoL	HF-4
4	Variance of Laplacian	VoL	LF
5	Spatial Frequency	SF	LF
6	Tenengrad	TG	HF-1
7	Brenner's Measure	BM	LF

After the selection of seven relevant features, the Region of Interest (ROI) of liver samples corresponding to normal, HCC and HEM classes are trained and classified using an SVM classifier with 5-fold cross validation. In this approach, the entire training set images are randomly grouped into five equal sets. Of the five sets, four sets are selected for training and the remaining one set is chosen for validation. Likewise, the process is repeated five times so that every subset will be used as a validation set once. Totally 280 liver CT images as mentioned in table III comprising of healthy liver, HCC and HEM are considered for the task of training the classifier so as to derive the required classifier parameters. The classifier performance is evaluated on a set of 120 images chosen from different patients. The training and test set images are carefully chosen to be disjoint in nature. The details of the count of three classes of image samples chosen for training and testing the classifier are listed in table III.

To accentuate the significance of feature extraction and selection schemes used in this work, the classifier performance is compared individually with that of focus features in spatial domain and in NSCT domain without feature selection. Also the results obtained out of this study is compared with the state-of-the art texture feature namely, GLCM features in spatial and NSCT domain. Additionally, the results are compared with two recent research work on classification using GLCM features using Curvelet transform [3] and dual tree complex wavelet

transform (DTCWT) [6]. The confusion matrix for the proposed CAD system and the classifier performance derived from the confusion matrix for the test data test are presented in tables IV and V respectively. The performance assessment for the proposed CAD system and its comparison were done in terms of classification accuracy and Kappa's coefficient.

TABLE III. DATA SET COUNT

Class	No. of training set images	No of test set images
Normal	97	40
HCC	95	40
HEM	88	40
Total	280	120

TABLE IV. CONFUSION MATRIX FOR CLASSIFICATION WITH PROPOSED FEATURE EXTRACTION AND SELECTION SCHEMES

	Normal	HCC	HEM
Normal	39	0	1
HCC	0	40	0
HEM	2	0	38

The classifier performance is usually analyzed with the help of two metrics namely classification accuracy and kappa coefficient. The accuracy is a statistical metric that deals with the percentage of correct predictions made. In the multi class classifier, it is not wise to rely solely upon the accuracy metric as it would produce high accuracy with poor recognition rate, if the data set is imbalanced. Hence, a metric that takes into account the degree of agreement among the different classes is necessary in assessing the performance of the classifier. Accordingly, Kappa coefficient [25] is used in addition to classification accuracy metric to overcome the limitations of the classifier due to accuracy paradox problems and to handle well the multi-class and imbalanced class problems. The Kappa coefficient is figured out mathematically by examining the random accuracy along with the actual accuracy of the classifier. Random accuracy is defined as the sum of the products of reference probability and result probability for each class.

$$Kappa = \frac{Actual\ Accuracy - Random\ Accuracy}{1 - Random\ Accuracy} \quad (17)$$

It is evident from the results presented in table V, that the classifier performs extraordinary well with focus features extracted from the ROI samples in NSCT domain combined with RRPC feature selection scheme.

IV. CONCLUSION

In this study, an alternate method of categorization of liver lesions using focus measures in NSCT domain is attempted and successfully demonstrated that these focus measures can also be used as features in machine learning based classification scenario in addition to their conventional usage as metrics to quantify the image clarity and sharpness. The fact that focus measures exhibit some sort of pixel inter dependency and considers local neighbourhood in their algorithmic computation, gives an expectation that it provides details on the textural pattern of the image. Comparison results presented in table V

TABLE V. COMPARISON OF CLASSIFIER PERFORMANCE METRICS

Metric	Focus Features			GLCM Features			
	In NSCT domain with RRPC feature selection	In NSCT domain without feature selection	In spatial domain	In NSCT domain	In Curvelet Domain [3]	In DTCWT Domain [6]	In spatial domain
Classification Accuracy	97.5%	81.66%	66.67%	82.06%	80.83%	78.33%	65.38%
Kappa's Coefficient	0.962	0.725	0.5	0.737	0.713	0.675	0.488

ascertains that CAD system derived using focus features can at least yield classification accuracy on par with the classic GLCM features or even better than that. Additionally, the features selected from the directional components of an image *viz.* HF coefficients of NSCT play a role in identifying inherent trait of the image or region of interest in pattern recognition problems. The proposed feature extraction technique in conjunction with feature selection method undoubtedly aids in achieving high classification rate without the need for complex enhancement of input samples and with simple classifier design. From the confusion matrix, it is observed that none of the HCC samples are misinterpreted as normal or HEM and similarly there is no false interpretation of normal and HEM samples as HCC. This will certainly reduce the hardship of the medical experts in making correct diagnosis and unhesitatingly, the proposed CAD system will serve as a second opinion for them in their diagnostic procedure. This study explores the effectiveness of focus features for liver CT images which can be further extended to other medical images to demonstrate its efficacy in pattern recognition.

REFERENCES

- [1] S. Roy, Y. Chi, J. Liu, S. K. Venkatesh and M. S. Brown, "Three-Dimensional Spatiotemporal Features for Fast Content-Based Retrieval of Focal Liver Lesions," in *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 11, pp. 2768-2778, Nov. 2014.
- [2] Gletsos, M., Mougiakakou, S. G., Matsopoulos, G. K., Nikita, K. S., Nikita, A. S., & Kelekis, D. (2003). A computer-aided diagnostic system to characterize CT focal liver lesions: Design and optimization of a neural network classifier. *IEEE Transaction on Information Technology in Biomedicine*, 7(3), 153–162.
- [3] Mala, K., Sadasivam, V., Alagappan, S., et al.: ‘Neural network based texture analysis of CT images for fatty and cirrhosis liver classification’, *Appl. Soft Comput. J.*, 2015, 32, pp. 80–86.
- [4] Chang, C.-C., Chen, H.-H., Chang, Y.-C., et al.: ‘Computer-aided diagnosis of liver tumors on computed tomography images’, *Comput. Methods Programs Biomed.*, 2017, 145, pp. 45–51
- [5] Bilello, M., Gokturk, S.B., Desser, T., et al.: ‘Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venousphase CT’, *Med. Phys.*, 2004, 31, (9), pp. 2584–2593
- [6] Peng Yang and Guowei Yang, Feature Extraction using Dual-Tree Complex Wavelet Transform and Gray Level Co-occurrence Matrix, Neurocomputing, vol 127, pp. 212-220, Jul 2012.
- [7] U Rajendra Acharya, U Raghavendra, Hamido Fujita, et al. Automated Characterization of Fatty Liver Disease and Cirrhosis Using Curvelet Transform and Entropy Features Extracted from Ultrasound Images, *Computers in Biology and Medicine*, vol 79, pp. 250-258, Oct 2016.
- [8] Amir Reza Sadri , Sepideh Azarianpour, Maryam Zekri, Mehmet Emre Celebi & Saeid Sadri, “WN-based approach to melanoma diagnosis from dermoscopy images” *IET Image Process.* 2017, Vol. 11 No. 7, pp. 475-482.
- [9] K. R. Krishnan1 , S. Radhakrishnan, “Hybrid approach to classification of focal and diffused liver disorders using ultrasound images with wavelets and texture features”, *IET Image Proc.*, vol. 11 no. 7, Jul 2017, pp. 530-538
- [10] Qi Zhang, Yang Xiao, Shuai Chen, Congzhi Wang, & Hairong Zheng, “Quantification of elastic heterogeneity using contourlet-based texture analysis in shear-wave elastography for breast tumor classification” *Ultrasound in Med. & Biol.*, pp. 588-600, Feb 2015.
- [11] Fatemeh Moayedi, Zohreh Azimifar,Reza Boostani,Serajodin Katebi, “Contourlet-based mammography mass classification using the SVM family”, *Computers in Biology and Medicine* vol.40, no. 2 Apr 2010, pp.373–383.
- [12] M.Dong, Z.Wang, C. Dong, X. Mu, and Y. Ma, “Classification of Region-of-Interest in Mammograms using Dual Contourlet Transform and Improved KNN”, *J. of Sens.*, vol 2017, Nov 2017.
- [13] F. Pak, H.R. Kanan, A. Alikhassi, “Breast cancer detection and classification in digital mammography based on non-subsampled contourlet transform (NSCT) and super resolution”, *Computer Methods and Programs in Biomedicine*, vol 122, no.2 Nov. 2015, pp.89-107.
- [14] Berbar MA. “Hybrid methods for feature extraction for breast masses classification”, *Egyptian Informatics J*, vol 19, no. 1, Mar 2017, pp. 63-73.
- [15] J.S.Leena Jasmine, S.Baskaran, S.Govardhan, “Non Subsampled Contourlet Transform based Classification of Microcalcification in Digital Mammograms”*Proc. Int. Conf. On Modelling Optimization and Computing ICMO’2012* pp.622-631.
- [16] Kovesi.P, “Phase congruency: a low-level image invariant,” *Psychological Research*, vol. 64, no. 2, pp. 136–148, 2000.
- [17] A, Oppenheim, J.Lim, “The importance of Phase in Signal”, *Proceedings of the IEEE* 69, pp.529-541, 1981.
- [18] Jin Xu, Bo Tang, Haibo He, & Hong Man, “Semisupervised Feature Selection Based on Relevance and Redundancy Criteria” *IEEE Trans. Neural Networks & Learning Systems*, Vol 28. No.9, Sep 2017, pp.1974-1984.
- [19] Said Pertuz, DomènecPuig, MiguelAngelGarcia, “Analysis of focus measure operators for shape-from-focus” *Pattern Recognition* Vol.46 2013 , pp.1415–1432
- [20] Nayar.S.K, “Shape from focus system”, *Computer Vision and Pattern Recognition, 1992. Proc. CVPR ’92*, pp. 3012 – 308, Jun 1992.
- [21] Mukul V. Shirvaikar, “An Optimal Measure for Camera Focus and Exposure” in *Proc. IEEE SSST 2004*
- [22] F. Helmli, S. Scherer, Adaptive shape from focus with an error estimation in light microscopy, in: Proceedings of the International Symposium on Image and Signal Processing and Analysis, 2001, pp. 188–193.
- [23] Wei Huang, Zhongliang Jing, “Evaluation of focus measures in multi-focus image fusion”, *Pattern Recognition Letters*, Vol. 28 , pp. 493 – 500, 2007.
- [24] J. Pech Pacheco, G. Cristobal, J. Chamorro Martinez, J. Fernandez Valdivia, Diatom autofocusing in brightfield microscopy: a comparative study, in: *Proc. of the Int. Conf. on Pattern Recognition*, vol. 3, 2000, pp. 314–317.
- [25] A. Thelen, S. Frey, S. Hirsch, P. Hering, Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood- size, and height value interpolation, *IEEE Transactions on Image Processing* 18 (2009) 151–157.
- [26] Helmut Küchenhoff , Thomas Augustin, Anne Kunz, “Partially identified prevalence estimation under misclassification using the kappa coefficient” *Int. J. of Approximate Reasoning*, Vol 53 (2012) pp.1168–1182.

Detection and Estimation of Multiple DoA Targets with Single Snapshot Measurements

Rakshith Jagannath

Department of Electrical Engineering
Indian Institute of Technology Madras
Email: ee13d005@ee.iitm.ac.in

Abstract—In this paper, we explore the problems of detecting the number of narrow-band, far-field targets and estimating their corresponding directions of arrivals (DoAs) from single snapshot measurements. We use the principles of sparse signal recovery (SSR) for detection and estimation of multiple targets. In the SSR framework, the DoA estimation problem is grid based and can be posed as the lasso optimization problem. The corresponding DoA detection problem reduces to estimating the optimal regularization parameter (τ) of the lasso problem for achieving the required probability of correct detection (P_c). We propose finite sample and asymptotic test statistics for detecting the number of sources with the required P_c at moderate to high signal to noise ratios. Once the number of sources are detected, or equivalently the optimal $\hat{\tau}$ is estimated, the corresponding DoAs can be estimated by solving the lasso with regularization parameter set to $\hat{\tau}$.

I. INTRODUCTION

Detection, estimation and tracking of targets are the primary functions of radar based localization systems. A main challenge frequently faced by these systems is the problem of limited measurements due to limited availability of sensors. In such cases, it is essential to exploit the sparsity of the targets in the array manifold (spatial domain) for the purpose of detection and estimation of the sources. In this work, we focus on the problems of detecting the number of narrow-band, far-field targets and estimating their corresponding direction of arrivals (DoAs) from single snapshot measurements.

The signal model used for detection and estimation in single snapshot DoA problem relates the observed measurements as a continuous and non-linear function of the DoA parameters [1]. As the DoA parameters are sparse in the spatial domain, sparse signal recovery (SSR) based techniques can be used for detection and estimation. In the SSR framework, the continuous DoA signal model can be approximated into three classes, namely on-grid, off-grid and grid-less [2]. In the on-grid SSR framework, the signal model for estimation is obtained by the discretization of the continuous array steering manifold over a selected interval of DoAs to construct the array steering matrix over an estimation grid of DoAs. The true DOA targets are then assumed to lie on the estimation grid. A number

of estimators have been proposed for DoA estimation. The SSR based estimators essentially use the lasso estimator in its various forms for estimation of the DoAs [3]. However, the lasso regularization parameter (τ), which controls the number of sources that are estimated is usually chosen empirically. In this paper, we deal only with the on-grid framework to explore the problem of finding the relationship between the lasso regularization parameter (τ) and the detection performance metrics like the probability of correct detection (P_c), the probability of miss-detection (P_m) and the probability of false alarm (P_f).

For the case of a single source in noise model, the regularization parameter estimate, $\hat{\tau} = \sigma\sqrt{-\ln(P_f)}$ for a given probability of false alarm P_f and noise variance σ , was obtained in [4] using the generalized likelihood ratio test (GLRT). However, for multiple targets, it is well-known that the GLRT selects the largest model [5]. Detection algorithms based on cross-validation and information criteria principles like Bayesian information criteria and minimum description length have been proposed in [6], [7]. However, these algorithms are known to suffer in detection performance for small number of snapshots and are mostly not even applicable for the single snapshot case [8]. Also, the relationship between τ and the probability of correct detection, P_c (or P_f) have not been obtained in these algorithms. A number of asymptotic (in M , the number of measurements) results which are the SSR counterparts to the martingale stability theorem (derived for maximum likelihood estimation framework) exist in the literature wherein the optimal regularization parameter ($\hat{\tau}$) is derived to minimize the lasso estimation error [9]. However, minimum estimation errors does not necessarily mean that sparsity and support of the estimate is same as the original parameter, which is required to control the P_c (or P_f) in the detection framework. In [10], the co-variance test statistics has been proposed for real measurements to obtain the optimal τ . The authors obtain an asymptotic cumulative distribution function (c.d.f) for the co-variance test statistics, which can then be used to obtain the optimal τ for given \tilde{P}_c , which is an approximation for P_c .

In this work, we use the on-grid DoA measurement model. In the moderate to high SNR regime, we propose Test-A, finite sample and asymptotic covariance tests for detection of multiple targets under the orthogonality

assumption on the model. Further, we also propose Test-D for detection of multiple target for the general on-grid model. We also derive the corresponding c.d.fs of the test statistics. Finally, we compare the performance of all these tests through simulations and discuss their merits.

Beamformers and sub-space based methods like ESPRIT and MUSIC cannot be used with single snapshot measurements for detecting multiple sources with adequate performance [11]. This is because all these techniques require the prior knowledge of the number of sources, which is unknown. Additionally, the sub-space based methods also require an estimate of measurement co-variance matrix, which requires multiple snapshots.

II. SIGNAL MODEL

We consider an array of M elements, impinged by an unknown number (S) of sources. The measurements at each element can be expressed as a superposition of S elementary waveforms ($a(\alpha_i, d)$), each containing unknown angles $\alpha_i \in [\kappa_1, \kappa_2]$ as,

$$\tilde{b}(d) = \sum_{i=1}^S s_i a(\alpha_i, d) + v(d),$$

where $v(d)$ is a white Gaussian noise process with zero mean and variance $2\sigma^2$, s_i are the weights and $\tilde{b}(d)$ are the measurements over the spatial variable $d = 1, 2, \dots, M$. The recovery problem now reduces to detecting the number of sources S , estimating their corresponding weights s_i and parameters α_i , which is non-linear [12].

In the grid based signal model for detection and estimation, the interval $[\kappa_1, \kappa_2]$ is discretized into N bins, each of size r to obtain the estimation grid, ρ_1, \dots, ρ_N . Let x_k denote the weight corresponding to the source in k^{th} bin. The discrete model approximation for $\tilde{b}(d)$ is [12],

$$b(d) = \sum_{k=1}^N x_k a(\rho_k, d) + v(d).$$

The above equation can be expressed in vector form as,

$$b(d) = \mathbf{a}^T(d)\mathbf{x} + v(d),$$

where $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$ and $\mathbf{a}(d) = [a(\rho_1, d), \dots, a(\rho_N, d)]^T$. Stacking the measurements, we obtain

$$\mathbf{b}_{M \times 1} = \mathbf{A}_{M \times N} \mathbf{x}_{N \times 1} + \mathbf{v}_{M \times 1}, \quad (1)$$

where \mathbf{b} is the measurement vector, $\mathbf{A} = [\mathbf{a}(0), \mathbf{a}(1), \dots, \mathbf{a}(M-1)]^T$ is the array steering matrix (with $M \leq N$) and \mathbf{x} is the signal of interest which is sparse.

Let $\boldsymbol{\alpha}$ be the vector representing S source locations (actual DoAs) and let $\hat{\boldsymbol{\rho}}$ represent the \hat{S} location estimates of the sources. We define the probability of correct detection (P_c) as the probability that all the sources and their locations are detected correctly, i.e. $P_c = \mathbb{P}\{\hat{\boldsymbol{\rho}} = \boldsymbol{\alpha}\}$, similarly the probability of miss (P_m) is defined as the probability that one or more sources is not detected, i.e. $P_m = \mathbb{P}\{\hat{S} < S, \hat{\rho}_i = \alpha_i, i = 1, 2, \dots, \hat{S}\}$ and the

probability of false alarm, $P_f = 1 - P_c - P_m$. We define the signal to noise ratio, SNR as $\mathbb{E}\{\|\mathbf{Ax}\|_2^2\}/\mathbb{E}\{\|\mathbf{v}\|_2^2\}$.

Problem description: Given the measurements (\mathbf{b}), the array steering matrix \mathbf{A} , SNR and the required probability of correct detection P_c . The goal is to propose test statistics to detect the number of sources (\hat{S}) and their corresponding locations ($\hat{\rho}_i$) on the estimation grid. The proposed tests should achieve the required probability of correct detection P_c .

III. DETECTION OF MULTIPLE TARGETS

In this section, we briefly review the lasso estimator, the lasso path and propose tests for joint detection and estimation of DoAs from single snapshot measurements.

The Lasso Estimator: The lasso estimator for the DoA model in (1) is given by the solution of the following optimization problem.

$$\hat{\mathbf{x}}(\tau) = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{b} - \mathbf{Ax}\|_2^2 + \tau \|\mathbf{x}\|_1, \quad (2)$$

where $\hat{\mathbf{x}}(\tau)$ is the estimate of \mathbf{x} and $\tau \in [0, \infty)$ is the regularization parameter which controls the sparsity of $\hat{\mathbf{x}}$. Applying KKT conditions to (2), the lasso solution can be characterized as follows,

Theorem 1. *For a certain value of τ , the solution to (2) is characterized by*

$$\mathbf{a}_i^H (\mathbf{b} - \mathbf{Ax}) = \tau \frac{\hat{x}_i}{|\hat{x}_i|} \quad \forall \hat{x}_i \neq 0, \quad (3)$$

$$|\mathbf{a}_i^H (\mathbf{b} - \mathbf{Ax})| < \tau \quad \forall \hat{x}_i = 0, \quad (4)$$

where \hat{x}_i , $i = 1, 2, \dots, N$ is the i^{th} entry of $\hat{\mathbf{x}}$ and \mathbf{a}_i is the i^{th} column of \mathbf{A} . The singular points (knot points) occur when the second condition is changed to $\tau = \max_{\{i|\hat{x}_i=0\}} |\mathbf{a}_i^H (\mathbf{b} - \mathbf{Ax})|$.

Proof. See [3]. □

We observe that the lasso solution for the special case of orthogonal array steering matrix ($\mathbf{A}^H \mathbf{A} = \mathbf{I}$) reduces to following thresholding estimator,

$$\hat{x}_j(\tau) = \begin{cases} \mathbf{a}_j^H \mathbf{b} - \tau \frac{\hat{x}_j}{|\hat{x}_j|} & \text{if } |\mathbf{a}_j^H \mathbf{b}| > \tau \\ 0 & \text{if } |\mathbf{a}_j^H \mathbf{b}| \leq \tau \end{cases} \quad (5)$$

We now discuss the behavior of $\hat{\mathbf{x}}$ for variations in τ which is called the lasso path. The lasso path can be obtained using the iterative algorithm described in [3].

Lasso Path: The lasso estimator $\hat{\mathbf{x}}(\tau)$ is a continuous and piecewise linear function of τ . The points τ_k with $\tau_1 \geq \dots \geq \tau_k \dots \geq \tau_r$, where the slope of the function $\hat{\mathbf{x}}(\tau)$ changes are called knots (or singular points) [3]. For all $\tau \geq \|\mathbf{A}^H \mathbf{b}\|_\infty$, the lasso estimate $\hat{\mathbf{x}}(\tau) = \mathbf{0}$. For decreasing τ , each knot τ_k marks the entry or removal of some variable from the current active set (J), which is the index set corresponding to non-zero entries of $\hat{\mathbf{x}}(\tau_{k-1})$. Hence, the active set remains constant in between the knots. For a matrix \mathbf{A} satisfying the special positive cone condition

(example orthogonal matrices), no variables are removed from the active set as τ decreases and hence there are always M knots in the lasso path.

We observe that the sparsity changes only at the knots. The estimation algorithm of [3] sequentially iterates over the knot points, $\tau_k, k = 1, 2, \dots, r$ and calculates $\hat{\mathbf{x}}(\tau_k)$. So, we propose tests at the knot points to obtain a stopping condition for the iterative algorithm as the lasso solution varies from $\hat{\mathbf{x}}(\tau_1)$ to $\hat{\mathbf{x}}(\tau_S)$. Once, the tests detect the number of sources \hat{S} or equivalently τ_S , the DoAs can then be estimated by solving lasso at $\tau = \tau_S$.

A. Orthogonal Models

Here we assume that the array steering matrix is orthogonal ($\mathbf{A}^H \mathbf{A} = \mathbf{I}$) and the sources lie on the estimation grid (perfect grid matching). These assumptions make the analysis of the test statistics simpler for evaluating thresholds. Specifically, the components of the lasso estimate, $\hat{\mathbf{x}}$ in (5) are independent. Although, this scenario is not practical as it occurs only for antennas with infinite apertures, the insights obtained here are helpful in proposing tests while working with non-orthogonal (over-sampled) models. We now describe the covariance test as follows.

Covariance Test: The covariance test statistics is defined at the knots of the lasso path. At the k^{th} knot, the covariance test statistics is defined as [10],

$$T_k = \frac{1}{\sigma^2} \left(\mathbf{b}^H \mathbf{A} \hat{\mathbf{x}}(\tau_{k+1}) - \mathbf{b}^H \mathbf{A}_J \tilde{\mathbf{x}}_J(\tau_{k+1}) \right), \quad (6)$$

where J is the active set just before τ_k , $\tilde{\mathbf{x}}(\tau_{k+1})$ is the solution of the lasso problem using only the active model \mathbf{A}_J (columns of \mathbf{A} belonging to J), with $\tau = \tau_{k+1}$, i.e.

$$\tilde{\mathbf{x}}_J(\tau_{k+1}) = \arg \min_{\mathbf{x} \in \mathbb{R}^{|J|}} \frac{1}{2} \|\mathbf{b} - \mathbf{A}_J \mathbf{x}_J\|_2^2 + \tau_{k+1} \|\mathbf{x}_J\|_1. \quad (7)$$

Intuitively, the covariance test statistics defined in (6) is a function of the difference between $\mathbf{A} \hat{\mathbf{x}}$ and $\mathbf{A}_J \tilde{\mathbf{x}}_J$, which represents the fitted values of the model by including and leaving out the next \hat{x}_j (corresponding to the knot at τ_{k+1}), respectively. For the case of orthogonal \mathbf{A} , it can be shown [10, Lemma 1] that the covariance test statistics reduces to

$$T_k = \tau_k (\tau_k - \tau_{k+1}) / \sigma^2, \quad k = 1, 2, \dots, M-1, \quad (8)$$

where, the M knots of the lasso estimator $\hat{\mathbf{x}}(\tau)$ are given by $[\mathcal{I}, \boldsymbol{\tau}] = \text{sort}(|\mathbf{A}^H \mathbf{b}|)$. The function $\text{sort}(\mathbf{u})$ sorts the entries of \mathbf{u} in the descending order, \mathcal{I} is the collection of the corresponding indices of $|\mathbf{A}^H \mathbf{b}|$ and $\boldsymbol{\tau}$ is the vector of M knot points.

Now, let the number of non zero entries in the actual parameter \mathbf{x} be S . We define B as the event that the S sources are added to the estimate $\hat{\mathbf{x}}$ at the first S knot points of the lasso path:

$$B = \left\{ \min_{j \in \tilde{T}} |\mathbf{a}_j^H \mathbf{b}| > \max_{j \notin \tilde{T}} |\mathbf{a}_j^H \mathbf{b}| \right\}, \quad (9)$$

where \tilde{T} is the support of the original parameter \mathbf{x} (columns of \mathbf{A} corresponding to non-zero entries of \mathbf{x}).

Remark-1: Event B is defined to ensure that S active parameters (S sources) are added to the estimate $\hat{\mathbf{x}}$ in the first S knots, then the test statistics at $S+1$ knot and beyond would depend only on the truly inactive variables (noise). The detection tests proposed below are conditioned on event B . Hence, $P(B) = 1$ is a necessary condition for the detection tests to provide rate control ($\hat{P}_c = P_c$). However, we show in Lemma 1 that event $P(B) \rightarrow 1$, whenever the power of the weakest source is large compared to the noise power or whenever the detection is performed in the moderate to high SNR regime [10, Theorem-1]. Hence, detection at moderate to high SNR is a sufficient condition for $P(B) \rightarrow 1$ and the tests to provide rate control for a given P_c .

Lemma 1. $P(B) \rightarrow 1$ at moderate to high SNRs. \square

Proof. See Appendix-VI-B \square

From the above discussions, we conclude that the stopping decision at $(S+1)$ th knot is necessary for providing rate control in the moderate to high SNR regime. This requires the evaluation of c.d.f of T_{S+1} conditional on event B , given by

Theorem 2. The c.d.f of T_{S+1} , conditional on event B is,

$$F_{T_{S+1}}(\eta) = 1 - n \int_0^\infty y e^{-(y^2/2)} \left(1 - e^{-\frac{(y-\eta/\sqrt{n})^2}{2}} \right)^{n-1} dy,$$

where $n = M - S$.

Proof. See Appendix-VI-C \square

Now, with the knowledge of the c.d.f of T_{S+1} conditional on event B , the problem of finding the number of sources S reduces to the following hypothesis testing problem.

$$\begin{aligned} H_o &= T_k \text{ is distributed as } F_{T_{S+1}}. \\ H_a &= T_k \text{ is not distributed as } F_{T_{S+1}}. \end{aligned}$$

The idea is to evaluate the test statistics at each knot in the increasing order (from τ_M to τ_1) and compare the value to a threshold, η . The first instance, where $T_k > \eta$ is the stopping point, because conditional on B , the stopping point corresponds to the knot τ_S , where all the sources have been added to the lasso estimate $\hat{\mathbf{x}}$. The threshold, η is obtained from the tail probability of the c.d.f of T_{S+1} by fixing the required probability of correct detection, P_c

$$P_c = \mathbb{P}\{T_k \leq \eta\} = F_{T_{S+1}}(\eta). \quad (10)$$

We observe that the c.d.f of the covariance test, though an exact (non-asymptotic) distribution, requires numerical integration for evaluating the threshold at each knot, hence making the test complicated. In [10], the asymptotic c.d.f of $T_k, k > S$, conditional on event B is derived for real measurement model. The extension to complex measurement model is given by the following theorem,

Theorem 3. Let the magnitude of the smallest nonzero entry of \mathbf{x} is large compared to σ . Then event B is satisfied, i.e. $\mathbb{P}(B) \rightarrow 1$ and furthermore, for each fixed $l \geq 0$

$$[T_{S+1}, T_{S+2}, \dots, T_{S+l}] \xrightarrow{d} \left[\text{Exp}(1), \text{Exp}\left(\frac{1}{2}\right), \dots, \text{Exp}\left(\frac{1}{l}\right) \right],$$

conditional on B , i.e. l^{th} statistics after S converges independently to exponential distribution with mean $1/l$.

Proof. See Appendix-VI-D \square

We observe that although the asymptotic distribution of T_{S+1} is tractable, it converges very slowly ($2 \log M$), hence offering lesser control in-terms of P_c . So we now propose another test which are both easy to evaluate and exact.

Test-A: We note that, if event B is satisfied and there are S sources, then $A_k = \frac{\tau_{S+k}}{\sigma}$, $k = 1, \dots, M - S$ are the order statistics of Rayleigh random variables. We define the Rayleigh test statistics as

$$A_k = \frac{\tau_{k+S}}{\sigma}. \quad (11)$$

As τ_{S+1} is the first knot point corresponding to noise, hence P_c can be controlled by accurately detecting A_1 . The threshold for controlling P_c requires the c.d.f of A_1 which is given by,

Theorem 4. The c.d.f of A_1 conditional on event B is,

$$F_{A_1}(x) = (1 - \exp(-x^2/2))^{M-S}. \quad (12)$$

Proof. A_1 is the maximum of the i.i.d Rayleigh random variables and hence its c.d.f is obtained by (12). \square

Similar to other tests, the problem of finding S sources reduces to comparing A_k with a threshold (η) at each knot point. The threshold is obtained from the c.d.f (12) by fixing F_{A_1} to the required P_c . We summarize the steps for detection and estimation of DoAs with orthogonal measurement model in Algorithm-1.

Algorithm 1 Algorithm for Detection and Estimation

- 1: **Inputs:** \mathbf{b} , \mathbf{A} , $\boldsymbol{\eta}$ (obtained by inverting the c.d.f).
 - 2: **Initialize:** Set $i = M - 1$, $\hat{S} = 0$, $[\mathcal{I}, \boldsymbol{\tau}] = \text{sort}(|\mathbf{A}^H \mathbf{b}|)$.
 - 3: **Evaluate:** Evaluate the test statistics A_i .
 - 4: **Decision:** If $A_i \geq \eta_i$ go to step 6
 - 5: **Iterate:** Decrease i by 1 and iterate from step 3.
 - 6: **Outputs:** $\hat{S} = i$, $\hat{\mathcal{I}} = \mathcal{I}(1, 2, \dots, \hat{S})$, $\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}(\hat{S})$, $\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}(\hat{\mathcal{I}})$.
-

B. Non Orthogonal Models

We now obtain tests for the case where the estimation grid is over-sampled to $N >> M$ bins to obtain a fat array steering matrix (\mathbf{A}) and all the source locations are perfectly matched to the estimation grid. From the discussions on orthogonal models, we observed that test statistics to control P_c can be proposed at knot points. Hence, we will first study the knot points of the lasso for a fat matrix \mathbf{A} . The first knot point of the lasso occurs

at $\tau_1 = \max_k |\mathbf{a}_k^H \mathbf{b}|$. The process of finding the subsequent knots is summarized by the following iterative procedure.

- The active set $J = \{j_1, j_2, \dots, j_n\}$ is determined by solving (3) at τ_k .
- For each $k \notin J$, solve the following system of equations for a vector $\hat{\mathbf{x}} = [\hat{x}_1, \dots, \hat{x}_n]$ and a set Λ_j .

$$\left\{ \mathbf{a}_{j_l}^H (\mathbf{b} - \mathbf{A}_J \hat{\mathbf{x}}) = \Lambda_k \frac{\hat{x}_l}{|\hat{x}_l|} \right\}_{l=1}^n, |\mathbf{a}_j^H (\mathbf{b} - \mathbf{A}_J \hat{\mathbf{x}})| = \Lambda_k$$

If the system is infeasible, we put $\Lambda_j = 0$.

- The next singular point is given by, $\tau_{k+1} = \max_j \Lambda_j$.

We now propose a test at the knot points. The goal of the proposed test is to detect the $(S + 1)^{th}$ knot point (where S is unknown), conditional on event B .

Test-D: At the k^{th} knot, the D test statistics is,

$$D_k = \frac{\tau_{S+k}^2}{\sigma^2}. \quad (13)$$

Again, assuming event B is true (i.e., $P(B) \rightarrow 1$), we need to make a decision at $(S + 1)^{th}$ knot. Hence, we require the c.d.f of D_1 , given by

Theorem 5. The c.d.f of D_1 , conditional on event B is,

$$F_{D_1}(\eta) = \prod_{i=1}^{M-S} (1 - e^{-\eta/\varrho_i}), \quad (14)$$

where ϱ_i are the $M - S$ non-zero eigen values of the projection matrix \mathbf{Q}_{M-S} .

Proof. See Appendix-VI-A \square

Similar to other tests, the problem of finding S sources reduces to comparing D_k with a threshold (η) at each knot point. The threshold is obtained from the c.d.f (14) by fixing F_{D_1} to the required P_c .

IV. NUMERICAL SIMULATIONS

We now evaluate the performance of the proposed detection tests. The simulation setup consists of a uniform linear array (ULA) with 8 antennas, which is receiving signal from S equal power sources. The total source power is, $\mathbb{E}\{\|\mathbf{x}\|^2\} = 1$. We generate the estimation grid $\boldsymbol{\rho}$ by uniformly sampling the interval $[-\pi/2, \pi/2]$ into $N = 8$ bins for orthogonal models and $N = 16$ bins for over-sampled model. We use the orthogonal model for evaluating the performance of T_k and A_k and use the over-sampled model for evaluating the performance of D_k . The array steering matrix, \mathbf{A} of size $M \times N$ is generated as explained in section II. The sources are then detected by using the detection tests as described in Algorithm-1 for orthogonal models and in steps described in Section-III-B for over-sampled model. The threshold is set to maintain the $P_c = 0.99$ (or $P_e = P_m + P_f = 0.01$). In the following, we use Monte-Carlo simulations for $L = 100000$ noisy realizations to evaluate the performance. Figure-1 shows the variation of $P(B)$ w.r.t SNR and confirms

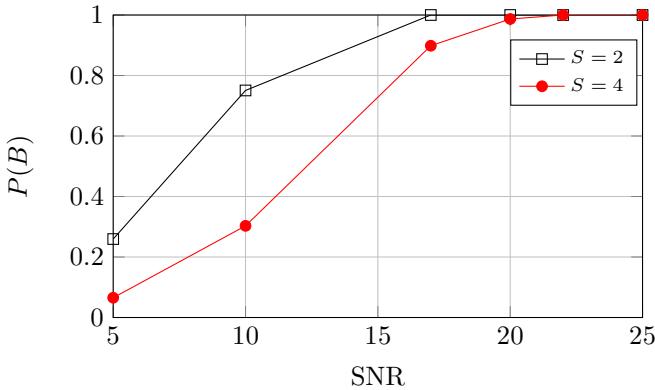


Fig. 1: $P(B)$ vs SNR for $S = 2$ and $S = 4$ source scenarios. $P(B)$ is calculated using knots of orthogonal model.

(a) \hat{P}_c obtained by the tests for different SNRs, $S = 2$, equal power

Statistic	T_k (Finite)	T_k (Asymp)	A_k	D_k
SNR = 10 dB	0.1155	0.2112	0.1432	0.0011
SNR = 15 dB	0.9225	0.9373	0.9562	0.5251
SNR = 20 dB	0.9908	0.9677	0.9906	0.9921
SNR = 25 dB	0.9906	0.9682	0.9902	0.9917
SNR = 50 dB	0.9902	0.9684	0.9900	0.9930

(b) \hat{P}_c obtained by the tests for different SNRs, $S = 4$, equal power

Statistic	T_k (Finite)	T_k (Asymp)	A_k	D_k
SNR = 10 dB	0.0011	0.0038	0.0010	0.0086
SNR = 15 dB	0.2122	0.3812	0.2681	0.4700
SNR = 20 dB	0.9890	0.9612	0.9903	0.9001
SNR = 25 dB	0.9895	0.9628	0.9901	0.9930
SNR = 50 dB	0.9903	0.9628	0.9901	0.9927

TABLE I: \hat{P}_c obtained by the proposed test statistics, T_k , A_k and D_k for a ULA with $M = 8$ antennas impinged by S sources. The threshold is obtained by setting $P_c = 0.99$.

our result in Lemma-1. Table-I shows \hat{P}_c obtained by the detection algorithms. The number of sources (S) received are indicated in the caption. From Table-I, we observe that none of the proposed tests achieve $\hat{P}_c = P_c$ at low SNRs as $P(B) < 1$. Specifically, for $\text{SNR} < 17$ dB in two source and $\text{SNR} < 22$ dB in the four source scenarios, $\hat{P}_c \neq P_c$. Next, we observe that all the finite sample tests (T_k (Finite), A_k and D_k) give perfect rate control ($\hat{P}_c = P_c$) at moderate to high SNRs where event B is true. Finally, we observe that the asymptotic covariance test (T_k (Asymp)) does not give rate control (i.e. $\hat{P}_c < P_c$) even at high SNR. From the observations, we can conclude that the proposed tests maintain rate control ($\hat{P}_c = P_c$) at moderate to high SNRs, where event B is true. We note that the evaluation of threshold (η) for the finite sample covariance test requires numerical integration, which makes it the most complex test, but there is no gain in-terms of rate control compared to Test-A. We also note that although the tests have been performed for $P_c = 0.99$, the rate control for higher P_c was also observed and upto 7 sources could be detected for orthogonal models. As explained in Section-I, other schemes for multiple target detection do not offer rate

control w.r.t P_c . Hence we do not compare our tests with other schemes.

V. CONCLUSIONS

In this work, we propose tests for jointly detecting and estimating multiple sources using single snapshot measurements at moderate to high SNR. These tests can also be interpreted as stopping criterion for homotopy based lasso estimators, since they provide a stopping threshold as the lasso estimator travels the lasso path. The proposed algorithms offer control over the probability of correct detection of the sources by choosing the appropriate threshold. Although we have applied the algorithm only for DoA problem, the algorithm can be used for any linear model with Gaussian noise problem. Achieving similar control over probability of correct detection in case of grid mismatch is an interesting problem for future work.

REFERENCES

- [1] H. L. Van Trees, "Optimum array processing (detection, estimation, and modulation theory, part IV)," *Wiley-Interscience*.
- [2] Z. Yang, J. Li, P. Stoica, and L. Xie, "Sparse Methods for Direction-of-Arrival Estimation," *ArXiv e-prints*, Sep. 2016.
- [3] A. Panahi and M. Viberg, "Fast candidate points selection in the lasso path," *IEEE Signal Processing Letters*, vol. 19, no. 2, pp. 79–82, Feb 2012.
- [4] J. Fuchs, "The generalized likelihood ratio test and the sparse representations approach," in *Image and Signal Processing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 6134, pp. 245–253.
- [5] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. New Jersey: Prentice-Hall Inc, 1993.
- [6] P. Boufounos, M. F. Duarte, and R. G. Baraniuk, "Sparse signal reconstruction from noisy compressive measurements using cross validation," in *IEEE/SP 14th Workshop on Statistical Signal Processing*, 2007. IEEE, 2007, pp. 299–303.
- [7] C. D. Austin, R. Moses, J. Ash, and E. Ertin, "On the relation between sparse reconstruction and parameter estimation with model order selection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 3, pp. 560–570, 2010.
- [8] B. M. Radich and K. M. Buckley, "Single-snapshot DOA estimation and source number detection," *IEEE Signal Processing Letters*, vol. 4, no. 4, pp. 109–111, April 1997.
- [9] A. Panahi and M. Viberg, "Maximum a posteriori based regularization parameter selection," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [10] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani, "A significance test for the lasso," *Annals of statistics*, vol. 42, no. 2, pp. 413–468, 2014.
- [11] P. Häcker and B. Yang, "Single snapshot DOA estimation," *Advances in Radio Science*, vol. 8, pp. 251–256, 2010.
- [12] C. Ekanadham, D. Tranchina, and E. Simoncelli, "Recovery of sparse translation-invariant signals with continuous basis pursuit," *IEEE Transactions on Signal Processing*, vol. 59, no. 10.
- [13] R. K. Mallik, "On multivariate rayleigh and exponential distributions," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1499–1515, June 2003.
- [14] B. Arnold, N. Balakrishnan, and H. Nagaraja, *A First Course in Order Statistics*, ser. Classics in Applied Mathematics. SIAM.
- [15] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction*, ser. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.
- [16] I. Weissman, "Estimation of parameters and large quantiles based on the k largest observations," *J. Amer. Statist. Assoc.*, vol. 73, no. 364, pp. 812–815, 1978.

VI. APPENDIX

A. Proof of Theorem-5

Assuming moderate to high SNR regime (event B is true) and S sources. Let $J = \{j_1, \dots, j_S\}$ be the active set, after S knot points. Now, at the $(S+1)^{st}$ knot point, $\tau_{S+1} = \max_{k \notin J} \Lambda_k$, $\Lambda_k = |\mathbf{a}_k^H(\mathbf{b} - \mathbf{A}_J \hat{\mathbf{x}}_J)|$ for some $k \in J^c$ and $\hat{\mathbf{x}}_J$ satisfies $\Lambda_k \mathbf{1} = |\mathbf{A}_J^H(\mathbf{b} - \mathbf{A}_J \hat{\mathbf{x}}_J)|$, $k \in J^c$. Hence, we obtain the following set of $|J|$ equations for $\hat{\mathbf{x}}$

$$|\mathbf{a}_k^H(\mathbf{b} - \mathbf{A}_J \hat{\mathbf{x}}_J)| = |\mathbf{a}_{j_i}^H(\mathbf{b} - \mathbf{A}_J \hat{\mathbf{x}}_J)| \quad \forall j_i \in J, k \in J^c. \quad (15)$$

Solving for $\hat{\mathbf{x}}$ from the above equations and substituting back in the expression for Λ_r , we obtain $\Lambda_r = |\mathbf{a}_r^H \mathbf{Q}_{M-S} \mathbf{v}|$, $r \in J^c$, where \mathbf{Q}_{M-S} is a projection matrix with S zero eigen values. Since, \mathbf{v} is a complex Gaussian random variable with zero mean and variance σ^2 , each Λ_r^2/σ^2 are correlated χ^2 random variables, hence the test D_1 is a maximum of correlated χ^2 random variables whose c.d.f is given by,

$$\begin{aligned} &= F_{D_1}(\eta) = \mathbb{P}(D_1 \leq \eta) = \mathbb{P}(\max_{r \in J^c} \Lambda_r^2/\sigma^2 \leq \eta), \\ &= \mathbb{P}(\Lambda_1^2/\sigma^2 \leq \eta, \dots, \Lambda_{|J|^c}^2/\sigma^2 \leq \eta) = \int_0^\eta f_{\mathbf{u}}(\mathbf{u}) d(\mathbf{u}), \\ &\stackrel{(a)}{=} \int_0^\infty f_{\mathbf{u}}(\mathbf{u}) \mathbb{I}(\mathbf{u}, \eta) d(\mathbf{u}) \stackrel{(b)}{=} \int_0^\infty \hat{f}_{\mathbf{z}}(\mathbf{z}) \prod_{i=1}^{M-S} \frac{(1 - e^{-j\eta z_i})}{j z_i} d\mathbf{z}, \\ &= \int_0^\infty (\det(\mathbf{I} - j \text{Diag}(\mathbf{z}) \mathbf{Q}_{M-S}))^{-1} \prod_{i=1}^{M-S} \frac{(1 - e^{-j\eta z_i})}{j z_i} d\mathbf{z}, \\ &= \prod_{i=1}^{M-S} \int_{z_i=0}^\infty \frac{(1 - e^{-j\eta z_i})}{j z_i (1 - j \varrho_i z_i)} dz_i = \prod_{i=1}^{M-S} (1 - e^{-\eta/\varrho_i}). \end{aligned}$$

In the above equations $\mathbb{I}(\mathbf{u}, \eta)$ denotes a unit box from 0 to η , $f_{\mathbf{u}}$ denotes the joint p.d.f of $\Lambda_r, r \in J^c$ in (a) and is degenerate because \mathbf{Q}_{M-S} is singular, hence we use the Parseval theorem in (b) and the characteristic function of correlated χ^2 random variables [13] to evaluate the c.d.f.

B. Proof of Lemma-1

Here we show that event $\mathbb{P}(B) \rightarrow 1$ in the moderate to high SNR regime (when $\theta = \min_{j \in T} x_j \gg \sigma$). We choose ϵ s.t. $\epsilon \gg \sigma$ and $\theta \gg \epsilon$. Now, the knots $\tau_k, k = 1, 2, \dots, S$ are independent Rician random variables. Hence,

$$\mathbb{P}\left(\min_{k \in \tilde{T}} \tau_k \geq \epsilon\right) = \prod_{k=1}^S \mathbb{P}\left(\tau_k \geq \epsilon\right) \geq \prod_{k=1}^S \mathcal{Q}_1\left(\frac{\theta}{\sigma}, \frac{\eta}{\sigma}\right)$$

Where, $\mathcal{Q}_1\left(\frac{\theta}{\sigma}, \frac{\epsilon}{\sigma}\right)$ is the Macrum Q function, which tends to 1 as $\frac{\theta}{\epsilon}$ tends to infinity. Hence $\mathbb{P}\left(\min_{k \in \tilde{T}} \tau_k \geq \eta\right) \rightarrow 1$ for large $\frac{\theta}{\eta}$. Also simultaneously, we note that $\tau_k, k = S+1, S+2, \dots, M$ are i.i.d. Rayleigh random variables, hence $\mathbb{P}\left(\max_{k \notin \tilde{T}} \tau_k \leq \eta\right) = (1 - \exp(-\frac{\eta^2}{2\sigma^2}))^{M-S}$ which tends to

1 as $\frac{\eta}{\sigma} \rightarrow \infty$. Hence, $\mathbb{P}\left(\max_{k \notin \tilde{T}} \tau_k \leq \eta\right) \rightarrow 1$ for large $\frac{\eta}{\sigma}$. So, we can conclude that $\mathbb{P}(B) \rightarrow 1$ for large $\frac{\theta}{\sigma}$.

C. Proof of Theorem-2

In the moderate SNR regime, $\frac{\tau_j}{\sigma}, j = S+1, S+2, \dots, M$ are the order statistics of Rayleigh random variable with p.d.f $f(x)$ and c.d.f $F(x) = 1 - \exp(-x^2/2)$. Defining $M - S = n$ and $V_j = \tau_{S+j}/\sigma$, we have $V_n \leq \dots \leq V_j \leq \dots \leq V_1$. Defining $V_j = X_{n+1-i}$, we have $X_1 \leq \dots \leq X_i \leq \dots \leq X_N$.

We first require the joint pdf of V_1, V_2 or X_n, X_{n-1} . The joint pdf of consecutive order statistics is [14, Chapter-2]

$$f_{X_k, X_{k+1}}(x, y) = C_0 \{F(x)\}^{k-1} \{1 - F(y)\}^{n-k-1} f(x) f(y),$$

where $C_0 = \frac{n!}{(k-1)!(n-k-1)!}$. Substituting $k = n-1$,

$$f_{X_{n-1}, X_n}(x, y) = C \{F(x)\}^{n-2} f(x) f(y), \quad 0 < x < y < \infty,$$

where $C = \frac{n!}{(n-2)!}$. The joint pdf of X_n and $w = X_n - X_{n-1}$ is,

$$f_{w, X_n}(w, y) = C \{F(y-w)\}^{n-2} f(y-w) f(y), \quad 0 < w < y < \infty.$$

Now, the joint p.d.f of X_n and $T_{S+1} = X_n W$ is,

$$f_{T_{S+1}, X_n}(t, y) = C \{F(y-t/y)\}^{n-2} f(y-t/y) f(y) \frac{1}{y}, \quad 0 < t < y^2 < \infty.$$

Finally the p.d.f of T_{S+1} is obtained by integration of the above equation w.r.t. y . Hence,

$$f_{T_{S+1}}(t) = \int_{\sqrt{t}}^{\infty} C \{F(y-t/y)\}^{n-2} f(y-t/y) f(y) \frac{1}{y} dy.$$

Now the cdf of the co-variance test statistics is,

$$\begin{aligned} F_{T_{S+1}}(\eta) &= \int_0^\eta \int_{\sqrt{t}}^{\infty} C \{F(y-t/y)\}^{n-2} f(y-t/y) f(y) \frac{1}{y} dy dt, \\ &= 1 - n \int_{\sqrt{\eta}}^{\infty} y \exp(-y^2/2) \{1 - \exp \frac{-(y-\eta/y)^2}{2}\}^{n-1} dy. \end{aligned}$$

D. Proof of Theorem-3

We note that Rayleigh random variables (V_i) satisfy the Von-Mises condition, Hence \exists constants $a_M = F^{-1}(1-1/M) = \sqrt{2 \log(M)}$ and $b_M = pF'(a_M) = \sqrt{2 \log(M)}$ s.t. $b_M(\frac{V_1}{\sigma} - a_M) \xrightarrow{d} -\log(E_0)$, where $-\log E_0$ has type I extreme value distribution [10], [15]. From [16], for any fixed $l \geq 1$, the random variables $W_0 = b_M(\frac{V_1}{\sigma} - a_M)$ and $W_i = b_M(\frac{V_i - V_{i+1}}{\sigma})$, $i = 1, \dots, l$ converge jointly as $(W_0, W_1, W_2, \dots, W_l) \xrightarrow{d} (\log G_0, E_1/1, E_2/2, \dots, E_l/l)$, where G_0, E_1, \dots, E_l are independent and G_0 is Gamma distributed with scale parameter 1 and shape parameter l , and E_1, \dots, E_l are standard exponentials. We have,

$$\begin{aligned} T_{S+k} &= \frac{V_k}{\sigma^2} (V_k - V_{k+1}) = \left(a_M + \frac{W_0}{b_M} + \sum_{j=k}^l \frac{W_j}{b_M} \right) \frac{W_k}{b_M}, \\ &= W_k + \frac{1}{2 \log(M)} \left(W_0 + \sum_{j=k}^l W_j \right) W_k. \end{aligned}$$

Hence T_{S+k} converges to W_k which converges to $\text{Exp}(1/k)$ as $M \rightarrow \infty$.

Signal design and detection algorithms for quick detection under false alarm rate constraints

Alamuru Pavan Kumar
Payload Data Management Group,
U. R. Rao Satellite Centre,
Bangalore

Vineeth B. S.
Department of Avionics,
Indian Institute of Space Science and Technology,
Thiruvananthapuram

Abstract—In this paper, we consider the design of sequential detection algorithms for the low delay detection of a finite duration transient change signal from noisy observations of the signal under a false alarm rate constraint. Such design problems are motivated by the need to detect explicit control signals with low delay. We propose five heuristic detection algorithms that include algorithms that directly estimate the start time of the signals. In contrast to prior work, we also consider the case where the transient change signal can be apriori designed so as to optimize the detection delay as well as false alarm rate. Using simulations and numerical studies, we compare the average delay and false alarm rate performance of the above algorithms for different choices of the transient change signals.

I. INTRODUCTION

In this paper we consider the design of transient change signals and sequential detection algorithms for the low delay detection of a transient change signal from noisy observations under a false alarm rate constraint. Such design problems are motivated by the need to detect explicit control signals with low delay; the control signals can be designed. We propose five heuristic detection algorithms that include algorithms that directly estimate the start time of the signals. Using simulations and numerical studies, we compare the average delay and false alarm rate performance of the above algorithms for different control signal designs.

We note that the problem of designing detection policies in this context is related to the classical problem of quickest change detection [7], [11]. Prior work in the quickest change detection field have assumed the change to be persistent, i.e., once the change occurs, the system will stay in the in-change state forever. The problem we consider in this paper is a variant of the quickest transient change detection problem considered by Premkumar et al. [8]. In the transient change framework they considered, the change occurs at random (Geometrically distributed) time and persists for a Geometrically distributed random duration. Detection policies were obtained from a partially observed Markov decision process (POMDP) formulation of the problem. The detection of a randomly arriving profile was considered by Guépié et al. [3]. In this paper, the change corresponds to the presence of the

control signal, which changes the statistics of the observations. In contrast to prior work in quickest change detection, we consider the problem where the change or control signal persists for deterministic finite time and the control signal can be designed. We also propose low complexity heuristic policies with performance close to that of the policies obtained from a similar POMDP formulation.

As a possible application we consider the following scenario in a cognitive radio (CR) setting, where the primary user (PU) owns the spectrum, and the secondary users (SUs) are opportunistic users of the spectrum. In the overlay spectrum sharing method [1], the secondary user (SU) should detect the PU’s transmission need and is required to vacate the spectrum whenever the PU starts using the channel. We consider a scheme wherein the PU can explicitly signal an impending transmission to the SUs by transmitting a short control signal on a common control channel [1]. The explicit signalling needs to be purposefully kept short so as to reduce the overhead in signalling. Our objective in this paper can be seen as the design of SU detection algorithms to detect this short control signal. We note that the SU should detect the short control signal with minimum detection delay. However, since the SU should not vacate the spectrum unnecessarily, the minimum delay detection should be done under a false alarm rate constraint at the SU. The persistent change quickest detection framework has been used in the CR setting in [4] and [5], to design PU activity detection policies that minimizes the detection delay subject to an upper bound constraint on the false alarm rate. However, for many applications, such as ours, it is more appropriate to consider the PU control signal detection problem as a quickest transient change detection where the change persists for a finite time. In this paper, we design sequential detection algorithms for the low delay detection of the explicit control signals while optimally trading off the control signal detection delay with the false alarm rate.

We note that although the model that we study in this paper has been motivated from the CR setting discussed above, it can be applied to quickest change detection problems in the fields of quality control, navigation system integrity monitoring, and seismic data processing [10].

The main contributions of our paper are: we consider a transient control signal detection problem where the control signal can be designed, propose heuristic policies for detecting

The second author is the corresponding author. Email address: vineethbs@gmail.com.

the signal, and we evaluate and compare the average delay false alarm rate tradeoff for these policies for different control signal choices. We discuss our system model and state the tradeoff problem in Section II. We then formulate our problem as a Markov decision process in Section III. From the insights that we obtain in Section III we propose five heuristic policies in Section IV. Three heuristic policies are based on the quickest change detection framework, while two directly estimate the start time of the change. In Section V we evaluate the performance of the proposed heuristic policies.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We assume that the system evolves in discrete time, with slots indexed by $n \in \{0, 1, 2, \dots\}$. The signal is assumed to start at a random slot or epoch τ . As in the quickest detection framework the start slot τ is assumed to be Geometrically distributed¹ with success parameter p_c . The duration of the signal is assumed to be N slots, where N is finite and deterministic, so that the signal persists only for N slots and disappears after that. We note that there is an assumption of slot synchronism between the origin of the signal and the detector here; this needs to be addressed in future work. The phenomenon of signal existing for N slots can be represented using the Markov state transition model shown in Figure 1. At a given slot n , the signal's state is denoted as $S_n \in \mathbb{S}$, where \mathbb{S} is the state space consisting of $N + 2$ states: (i) pre-signal state denoted as 0, (ii) in-signal states $\{1, 2, \dots, N\}$, and (iii) post-signal state denoted as $N + 1$. Since the signal is assumed to start at the Geometrically distributed slot τ , we have that the transition probability from 0 to 1 is p_c . The other transition probabilities of the underlying Markovian state evolution S_n are as shown in Figure 1.

The change has to be detected by a detector or decision maker which does not observe the signal state S_n directly. The detector observes S_n partially through noisy observations; the observation in slot n is denoted as X_n . The cumulative distribution function (CDF) of X_n when $S_n = 0$ or $N + 1$ (i.e., when the signal is not present) is given by $F_0(\cdot)$, and that when $S_n = i, \forall i \in \{1, 2, \dots, N\}$ is given by $F_i(\cdot)$, which is assumed to be $\neq F_0(\cdot)$ for every $i \in \{1, 2, \dots, N\}$. We assume that the corresponding probability density functions (PDF) f_0 and $f_i \neq f_0$ exist for $i \in \{1, 2, \dots, N\}$. As an example, consider $f_0(x) = \mathcal{N}(x; 0, \sigma^2)$, i.e., Gaussian with mean 0 and variance σ^2 , and $f_i(x) = \mathcal{N}(x; \sqrt{P_i}, \sigma^2)$, Gaussian with mean $\sqrt{P_i}$ for all $i \in \{1, 2, \dots, N\}$. For our motivating example from CR, this models the case where if the PU is transmitting the signal and the received signal power level is P_i

¹We use the geometric distribution on τ to model an independent and identically distributed start time epoch. Similar models have been used in [11] and [8].

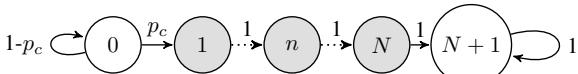


Fig. 1: Markov chain model for the change persisting for N slots

and P_i/σ^2 as the signal to noise ratio, otherwise the received signal power level is zero. We note that we are considering an extremely simple scenario, where unknown propagation losses are not accounted for. We assume that the observations X_n are conditionally independent given the state S_n , i.e., X_{n+1} is independent of X_n given S_n .

We note that the length of the signal N is under the designer's control. Furthermore, we also assume that F_i -s, $i \in \{1, 2, \dots, N\}$ can be partly controlled by the designer; for example the power level P_i can be chosen by the designer for every i . We also note that although the above example looks simplistic, the above model can be applied to more realistic wireless communication scenarios which include fading, under the assumption that the fade distribution is known. Then, the distributions $F_i(\cdot)$ and $F_0(\cdot)$ should model the observed value incorporating the distribution due to the random fading effect.

The detector can take either one of two actions in a slot n . In each slot n we can either: (0) stop and declare that the signal has been detected, or (1) continue sensing for the next slot $n+1$. The action in slot n is therefore denoted by $A_n \in \{0, 1\}$. We note that if the action is to continue sensing for the slot $n+1$, then the detector receives the observation X_{n+1} . A detection policy μ is a sequence $(A_0, A_1, \dots, A_T = 0)$ where T indicates the slot in which the control signal is declared to be detected. We note that T can be infinity, which is the case where the policy fails to detect the change (in our model the change happens with probability 1 since the change time is geometric). For a policy μ , we define the average detection delay (ADD) for μ as $ADD(\mu) = \mathbb{E}[\max(0, T - \tau)]$. For the policy μ , the probability of false alarm (P_{FA}) is defined as $P_{FA}(\mu) = \mathbb{P}\{S_T = 0\} = \mathbb{E}[\mathbb{I}\{T < \tau\}]$, which is the probability that the state in slot T , at which we have declared a change, is 0. We first consider the problem of designing sequential detection algorithms where our objective is to obtain a policy μ which minimizes $ADD(\mu)$ under an upper bound constraints on $P_{FA}(\mu)$ for a given set of N , $F_0(\cdot)$, and $F_i(\cdot), i \neq 0$. From [8], we have that such a policy can be obtained by considering the problem:

$$\underset{\mu}{\text{minimize}} \quad ADD(\mu) + \lambda P_{FA}(\mu). \quad (1)$$

where we define λ as a non-negative Lagrange multiplier. We note that the solution to this problem leads us to the characterization of a tradeoff between ADD and P_{FA} . The tradeoff is obtained by varying the parameter λ , giving more importance to P_{FA} as λ is increased. We note that the probability of detection is the fraction of time that $T < \infty$. If the average detection delay is finite, then that implies that the probability of detection is 1. Since the average detection delay partially incorporates the probability of detection we consider the $ADD - P_{FA}$ tradeoff rather than directly consider the detection probability. In the next section, we formulate the above problem as a completely observable Markov decision process (COMDP) in order to obtain insights for designing detection policies. Then we make our first steps towards the study of the signal design problem; we propose heuristic

policies using the COMDP framework and then we study the $ADD - P_{FA}$ tradeoff for several choices of N and $F_i(\cdot), \forall i \in \{1, 2, \dots, N\}$ in order to obtain the best tradeoff.

III. A COMDP FORMULATION FOR (1)

We note that since S_n is not directly observed by the detector, a policy μ which achieves the minimum in (1) could be very general, in that it prescribes an action A_n at slot n which can depend on the prior information about the system (such as p_c , the CDFs F_0 and F_i -s), and the observations X_1, \dots, X_n upto n . However, it is standard to show that [2, Chapter 5], there exists an optimal policy μ which prescribes an action A_n which is a function only of the *belief* vector \bar{p}_n . Here the belief vector $\bar{p}_n = (p_n(0), p_n(1), \dots, p_n(N), p_n(N+1))$ where $p_n(i) = \mathbb{P}\{S_n = i | X_1, X_2, \dots, X_n\}$ (the posterior probability of the state being i given all the observations upto slot n), and $\sum_{i=0}^{N+1} p_n(i) = 1$. We assume that $\bar{p}_0 = \bar{\nu}$ is given, where $\bar{\nu}$ is the initial belief that the detector has about the hidden state S_n . We note that $\bar{p}_{n+1}(i)$ is obtained from \bar{p}_n after receiving the observation X_{n+1} by Bayes' rule, if at slot n the action taken is to continue. Using Bayes' rule and the transition probabilities from Figure 1, we have that

$$p_{n+1}(i) \propto \begin{cases} (1-p_c)p_n(0)f_0(X_{n+1}), & \text{if } i = 0 \\ p_c p_n(0)f_1(X_{n+1}), & \text{if } i = 1 \\ p_n(i-1)f_i(X_{n+1}), & \text{if } i \in \{2, \dots, N\} \\ (p_N(n) + p_{N+1}(n))f_0(X_{n+1}), & \text{if } i = N+1 \end{cases} \quad (2)$$

We require that $\sum_{i=0}^{N+1} p_{n+1}(i) = 1$. We denote the above update operation which updates the belief vector \bar{p}_n to \bar{p}_{n+1} on deciding to continue and receiving the observation X_{n+1} as $U(\bar{p}_n, X_{n+1})$. If the action or decision is to stop, then we assume that \bar{p}_n transitions to a special absorbing state ϕ with probability 1. The optimization problem (1) can be viewed as a COMDP where the state space of the COMDP is the state space $[0, 1]^{N+1}$ of the belief vector², the state is the belief vector \bar{p}_n and the state transition is given by $U(\cdot, \cdot)$. The single stage cost $g(\bar{p}_n, a_n)$ for the COMDP as a function of the belief state \bar{p}_n and the action $a_n \in \{0, 1\}$ at slot n is $g(\bar{p}_n, a_n) = (1 - p_n(0))\mathbb{I}\{a_n = 1\} + \lambda p_n(0)\mathbb{I}\{a_n = 0\}$. We also assume that the single stage cost for the state ϕ for any action is 0. We note that then (1) is equivalent to minimizing $\mathbb{E}[\sum_{n=0}^{\infty} g(\bar{p}_n, A_n) | \bar{p}_0 = \bar{\nu}]$, which is a total cost infinite horizon COMDP with a zero-cost absorbing state. For such a COMDP, Bellman's optimality equation is:

$$J(\bar{p}) = \min [\lambda p(0), (1 - p(0)) + \mathbb{E}J(\bar{P})], \quad (3)$$

where if the minimum is taken over the actions 0 and 1 respectively. We note that $\bar{P} = U(\bar{p}, X)$ is the updated belief on observing the random observation X and $J(\bar{p})$ is the optimal total cost incurred starting from an initial belief of \bar{p} . We note that $\lambda p(0)$ represents the single stage cost for action 0, since the state transitions to ϕ which has zero cost. We note the optimal control policy μ for (1) can be obtained as an

²We need only $N+1$ components since $\sum_{i=0}^{N+1} p_n(i) = 1$

optimal function $\mu(\bar{p})$, where $\mu(\bar{p})$ is the minimizing action from the above optimality equation for a belief state \bar{p} . We note that an optimal policy obtained as above is independent of the initial state $\bar{p}_0 = \bar{\nu}$. For the optimal control policy μ (in fact for any control policy which is specified in the form of a function $r : [0, 1]^{N+1} \rightarrow A$), the detection process is as follows. We initialize $\bar{p}(0) = \bar{\nu}$. In each slot, we choose $A_n = \mu(\bar{p}_n)$. (or $r(\bar{p}_n)$). If $A_n = 1$, we observe X_{n+1} and update \bar{p}_n to \bar{p}_{n+1} .

However, since $J(\bar{p})$ is not known it is not possible to obtain the optimal control policy from (3) directly. Furthermore, since the state space of the COMDP is continuous, numerical algorithms such as value iteration cannot be directly applied. The insight that we obtain from the COMDP formulation is that heuristic policies can be defined as a function of the belief vector \bar{p}_n . In the next section, we define heuristic policies which tradeoff the detection delay with the probability of false alarm which are functions of the belief vector \bar{p}_n or its variants.

IV. HEURISTIC POLICIES

In this section, we propose five heuristic policies for the tradeoff problem in (1). All policies prescribe the action A_n as a function of belief vectors. We also propose two heuristic policies that directly estimate τ and thus detect the start of the signal using definitions of an approximate belief vector, which is different from the \bar{p}_n described previously.

(1) *Belief on signal state (BSS) policy*: The BSS policy is defined using the belief vector \bar{p}_n as defined in Section III. In every slot $n+1$ on receiving the observation X_{n+1} , the belief vector is updated from \bar{p}_n to \bar{p}_{n+1} using $U(\bar{p}_n, X_{n+1})$. We denote by $\sum_{i \neq 0} \bar{p}_n(i)$ by $b_{n,\text{change}}$, the belief that a change has happened by slot n . The BSS policy is parameterized by a threshold parameter t_{bcs} . In slot n , with belief vector \bar{p}_n we choose $A_n = 0$ if $b_{n,\text{change}} > t_{bcs}$ and $A_n = 1$ if $b_{n,\text{change}} \leq t_{bcs}$.

We now propose two policies which are obtained by approximating the state evolution itself, but which can be applied for the case of a signal for which $F_i(\cdot) = F_1(\cdot), \forall i \in \{1, 2, \dots, N\}$.

(2) *Policy from geometric change duration assumption (GCDA)*: We note that the N in-change states $1, 2, \dots, N$ were needed in the state model (Figure 1) since the change duration is deterministic. If N was geometrically distributed instead of being deterministic, then the change state is described by a 3-state Markov chain, wherein the N in-change states $1, 2, \dots, N$ are coalesced into a single in-change state 1 (as in [8]). The transition probabilities from state 1 to itself would be $1 - p_d$ and to state 2 is p_d , where p_d is the parameter of the geometric distribution on N . We consider an approximation where we model the change state as above. Since N is deterministic in our actual model, we choose $p_d = \frac{1}{N}$ in our approximation. We define a belief vector $\bar{p}_n = (p_n(0), p_n(1), p_n(2))$ such that $\sum_{i=0}^{N+1} p_n(i) = 1$. The belief vector can be updated using Bayes' rule as in [8] (we note that this update introduces a model error, since the actual

system does not evolve in the same way as assumed in the model). With $b_{n,\text{change}} = p_n(1) + p_n(2)$ we choose $A_n = 0$ if $b_{n,\text{change}} > t_{\text{gcda}}$ and $A_n = 1$ if $b_{n,\text{change}} \leq t_{\text{gcda}}$, where t_{gcda} is a threshold parameter.

(3) *Policy from persistent change assumption (PPA):* For defining the PPA policy, we consider an approximate state model which assumes that the signal is persistent. The underlying state can be denoted by 0 and 1 representing whether the signal is not present or present respectively. In this case, for the PPA policy we only need to maintain a belief value $p_n(1) = 1 - p_n(0)$ which is the posterior probability $\mathbb{P}\{S_n = 1 | X_1, \dots, X_n\}$. The belief is updated after each observation is received according to the Bayesian update rule as in [11, Section III, eq (14)]. For the PPA policy we have a threshold t_{ppa} and we choose $A_n = 0$ if $p_n(1) > t_{\text{ppa}}$ and $A_n = 1$ if $p_n(1) \leq t_{\text{ppa}}$.

We note that both the GCDA and PPA policies operates under model errors. We note that BSS, GCDA, and PPA policies using the initial belief $\bar{\nu}$ are able to take those cases into consideration where the change may have started before time 0 (e.g., if $\nu(2) = 1$). However, in many cases it is reasonable to consider the case where the change has not started at time 0 so that $\nu(0) = 1$. Under this assumption, we propose the following policies, which directly estimate the change point and could be useful in other scenarios also.

(4) *Posterior on change point τ (POT) policy:* Under the above assumption for POT policy we directly estimate the change point τ by using the posterior probability of τ . We denote by $q_n(m) = \mathbb{P}\{\tau = m | X_1^n\}$ for $m = \{0, 1, 2, \dots, \tau_{\max}\}$ and $X_1^n = (X_1, X_2, \dots, X_n)$. The support of τ is the whole of non-negative integers, but for implementing the policy we consider a truncated support³ where we assume that $\tau \in \{0, 1, \dots, \tau_{\max} = \lceil \frac{30}{p_c} \rceil\}$. For every observation X_{n+1} we receive, the posterior $q_n(m)$ is updated according to the following update rule:

$$q_{n+1}(m) \propto \begin{cases} q_n(m)f_{n-m+1}(X_{n+1}), \\ \text{if } m \leq n+1 \leq m+N-1, \text{ and} \\ q_n(m)f_0(X_{n+1}), \text{otherwise.} \end{cases}$$

We note that $\sum_{m=0}^{\tau_{\max}} q_{n+1}(m)$ is required to be 1. We define the posterior probability that the change has happened before n as $b_{\text{change},n} = \sum_{m=0}^{n-1} q_n(m)$. The POT policy is parameterized by a threshold t_{pot} , and we choose $A_n = 0$ if $b_{\text{change},n} > t_{\text{pot}}$ and $A_n = 1$ if $b_{\text{change},n} \leq t_{\text{pot}}$.

(5) *Belief on time after change (BTC) policy:* The BTC policy uses an extended state space (as shown in Figure 2) instead of the state model in Figure 1. The state (h, n) denotes that the current slot is n and that the change has happened h slots in the past. Then we have that the difference $n-h$ is the slot at which the change has happened. We note that at slot n' the system can only be in the states (h, n') . For specifying

³The truncated support using the upper limit τ_{\max} is used because of finite memory constraints. We choose the parameter τ_{\max} so that the probability that τ exceeds τ_{\max} is small. We choose $\tau_{\max} = \lceil \frac{30}{p_c} \rceil$, where the probability that $\tau > \tau_{\max}$ for $p_c=0.1$ is approximately 2×10^{-14} .

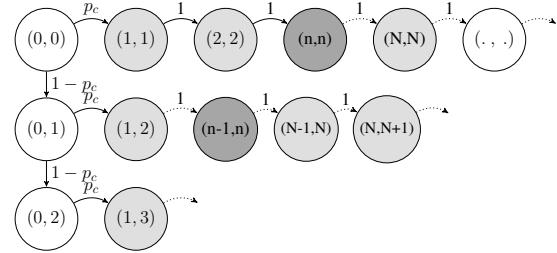


Fig. 2: State model corresponding to the BTC policy

this policy, we therefore have to maintain a belief vector $\bar{p}_n = (q_n(0), \dots, q_n(h), \dots, q_n(n))$. We update the beliefs according to (4).

$$q_{n+1}(h) \propto \begin{cases} (1 - p_c)q_n(0)f_0(X_{n+1}) & \text{if } h = 0, \\ p_c q_n(0)f_1(X_{n+1}) & \text{if } h = 1, \\ q_n(h-1)f_h(X_{n+1}) & \text{if } 2 \leq h \leq N, \\ q_n(h-1)f_0(X_{n+1}) & \text{if } h > N. \end{cases} \quad (4)$$

We define $b_{n,\text{change}} = \sum_{h=1}^n q_n(h) = 1 - q_n(0)$. The BTC policy is parameterized by a threshold t_{btc} , and we choose $A_n = 0$ if $b_{\text{change},n} > t_{\text{btc}}$ and $A_n = 1$ if $b_{\text{change},n} \leq t_{\text{btc}}$.

In terms of storage and computational complexity, the GCDA and PPA policies need only maintain 2 and 1 belief values, while the POT policies maintains τ_{\max} beliefs, BSS policy $N+1$ beliefs, and for the BTC policy, the dimension of the belief state vector increases linearly with time.

V. PERFORMANCE COMPARISON

In this section, we set $f_i(\cdot) = f_1(\cdot)$, $\forall i \in \{1, 2, \dots, N\}$, i.e., all the in-signal states have the same conditional distribution of observations. We first study the $ADD - P_{FA}$ tradeoff performance of the proposed policies for $f_0(\cdot)$ and $f_1(\cdot)$ being Gaussian distributed in Figures 3, 4, 5 and 6. We assume that $p_c = 0.1$ for the examples presented here. The thresholds for the policies ($t_{\text{bcs}}, t_{\text{gcda}}, t_{\text{ppa}}, t_{\text{pot}}$, and t_{btc}) are varied between 0.996 and 0.999999 in our experiments to obtain the tradeoff. We consider threshold values close to 1 to evaluate the low P_{FA} performance of detection policies. To compare how close the proposed policies perform to the optimal, we also plot a lower bound on the tradeoff curve which is obtained by solving the COMDP corresponding to the system with persistent change (or $N = \infty$). The COMDP is solved by discretizing the state interval $[0,1]$ to a set of points which are uniformly spaced with a distance of δ . The optimal policy is then obtained using the value iteration procedure [2]. We evaluate the tradeoff performance of the policies for different parameters P/σ^2 and N .

We have also compared the tradeoff performance of proposed policies with the correlation based non-sequential block policy [6] for the case of $f_0(\cdot)$ and $f_1(\cdot)$ being Gaussian. We have observed that in these cases the ADD is at least three to four times of the maximum ADD obtained by the GCDA, BSS, POT, or BTC policies; hence the tradeoff curves for the

block policy from [6] are not plotted in the figures in this section.

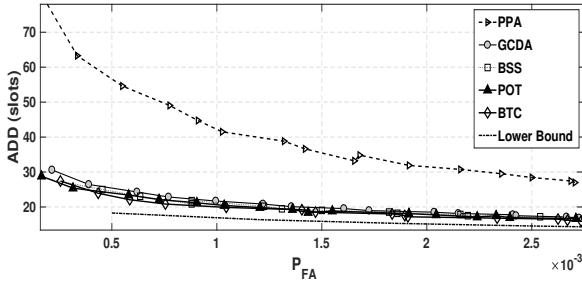


Fig. 3: The $ADD - P_{FA}$ tradeoff curves for the proposed policies for $P/\sigma^2 = -2.5$ dB, $N = 25$.

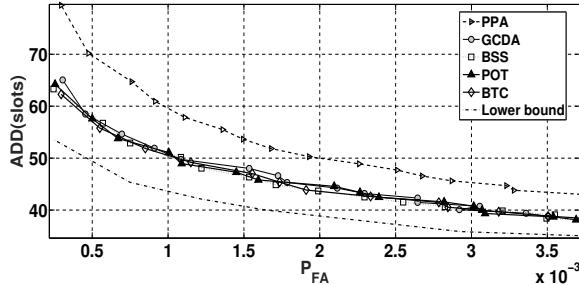


Fig. 4: The $ADD - P_{FA}$ tradeoff curves for the proposed policies for $P/\sigma^2 = -12$ dB, $N = 25$.

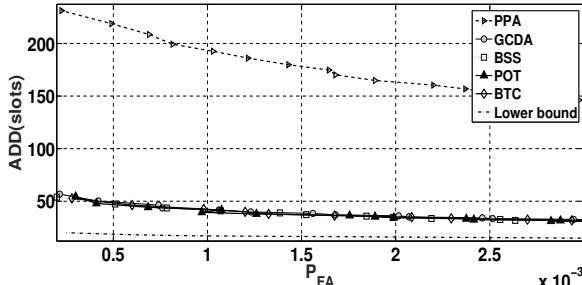


Fig. 5: The $ADD - P_{FA}$ tradeoff curves for the proposed policies for $P/\sigma^2 = -2.5$ dB, $N = 10$.

From Figures 3, 4, 5, and 6, our numerical results show that the BSS, GCDA, POT, and BTC policies have comparable performance. It is also intuitively satisfying to note that the minimum ADD for a given P_{FA} is a decreasing function of N . The GCDA and PPA policies are simple to describe and the GCDA policy has performance close to the optimal for the case of large P/σ^2 and N . In Figure 3, we observe that the ADD is less than the control signal duration $N = 25$ for low P_{FA} range which shows that the decision maker is able to detect the signal before N slots. However, the tradeoff performance of PPA is counterintuitive in the case of *small* N (e.g. $N = 10$ in Figures 5 and 6). In Figures 5 and 6, we observe that ADD for PPA is higher for $P/\sigma^2 = -2.5$ dB

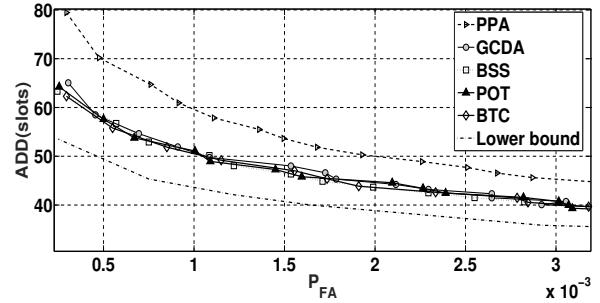


Fig. 6: The $ADD - P_{FA}$ tradeoff curves for the proposed policies for $P/\sigma^2 = -12$ dB, $N = 10$.

compared to that for $P/\sigma^2 = -12$ dB. This counterintuitive behaviour arises due to the persistent change model, under which a change would eventually happen with probability one. The model assumption that the change is persistent causes the belief probability $p_n(1)$ to drift upwards until the threshold is crossed, even though the likelihood ratio is 1 and therefore the drift caused due to the observation is zero. This is then registered as a change detection, even though it has happened due to the prior assumption in the model rather than due to the observations, leading to a smaller ADD .

We also consider an example where the observations X_n take discretized values. We model the distribution as follows. We assume that a symbol 0 is transmitted if the state is either 0 or $N + 1$, and a symbol 1 is transmitted when the state corresponds to a change. Due to noise and output discretization, we assume that the detector receives 1 when 0 is transmitted with probability ϵ and 0 when 1 is transmitted with probability $1 - \epsilon$. We consider the case where $\epsilon = 0.35$ and 0.45 in Figures 7 and 8 respectively. To compare how close the proposed policies perform to the optimal, we use the Perseus solver [9] to approximately solve the COMDP in Section III corresponding to (1). The GCDA, BSS, POT, and BTC policies have performance comparable to the policy obtained from Perseus solver for the discrete observations case.

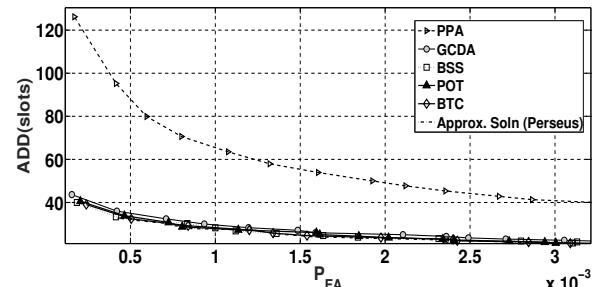


Fig. 7: The $ADD - P_{FA}$ tradeoff curves for the proposed policies for discrete observations, $N = 25$, $\epsilon = 0.35$.

We now consider the $ADD - P_{FA}$ tradeoff for some candidate signals. We compare the $ADD - P_{FA}$ tradeoff for the following three sequences of $F_i(\cdot)$ for $i \in \{1, 2, \dots, N\}$ for $N = 13$ and $N = 26$ in Figures 9 and 10:

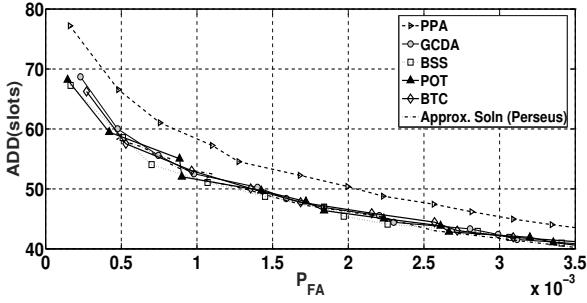


Fig. 8: The $ADD - P_{FA}$ tradeoff curves for the proposed policies for discrete observations, $N = 25$, $\epsilon = 0.45$.

- 1) (Ones): We consider $F_i(\cdot)$ to be Gaussian with a fixed mean P for every i .
- 2) (Barker): We consider $F_i(\cdot)$ to be Gaussian with the means μ_i chosen as Barker sequences of length 13 and 26. The signs of the length 13 mean sequence is $(+, +, +, +, +, -, -, +, +, -, +, -, +)$ with the magnitudes being P and the 26-length sequence is the 13-length repeated two times.
- 3) (IID): We consider $F_i(\cdot)$ to be Gaussian with the means μ_i chosen as a sample function of a Bernoulli IID random process of length 13 and 26. The signs of the 13 length mean sequence is $(+, -, +, -, +, -, -, -, +, -, +, -, +)$ and the 26 length sequence is $(-, -, +, -, -, -, +, -, +, -, +, -, +, -, +, -, -, -, +, +, +, -, +, +, +, +, -, +)$, with the magnitudes being P .

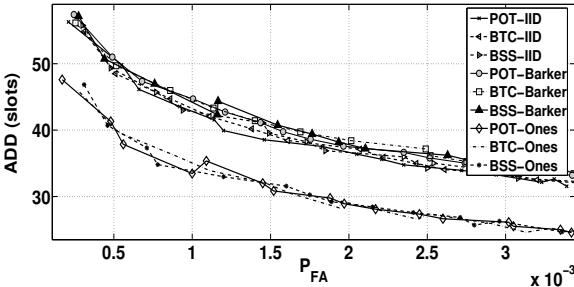


Fig. 9: The $ADD - P_{FA}$ tradeoff curves for the policies POT, BTC, and BSS for $P/\sigma^2 = -2.5$ dB, $N = 13$, for the three signals Ones, Barker, and IID sample

For the cases presented here, we obtain that the signal which has a constant mean has the best $ADD - P_{FA}$ tradeoff amongst the signals considered.

VI. CONCLUSION AND FUTURE WORK

In this paper, we considered the problem of designing detection policies for the quick detection of transient signals with applications to a CR context while trading off the false alarm rate. We proposed five heuristic policies: BSS, GCDA, PPA, POT, and BTC, of which POT and BTC are based on the direct estimation of the change time itself. The heuristic policies were motivated by a Markov decision theoretic formulation of

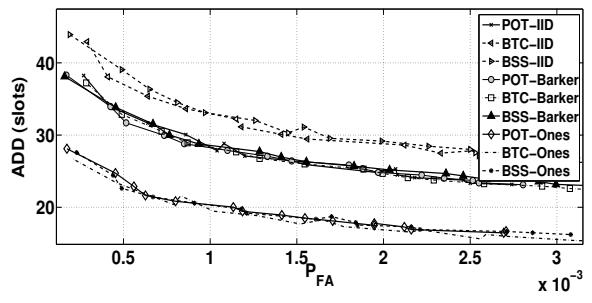


Fig. 10: The $ADD - P_{FA}$ tradeoff curves for the policies POT, BTC, and BSS for $P/\sigma^2 = -2.5$ dB, $N = 26$, for the three signals Ones, Barker, and IID sample

the detection problem. The $ADD - P_{FA}$ tradeoff performance for BSS, GCDA, POT, and BTC are shown to be similar using simulation and numerical studies for continuous and discrete valued channel models. We also observe that in the case of PPA, the persistent change assumption causes counterintuitive behaviour in certain cases, showing the effect of model-mismatch in such detection algorithm design problems. We also considered signal design; we evaluated the $ADD - P_{FA}$ tradeoff for candidate signals and obtained that the signal consisting of a constant mean has the best $ADD - P_{FA}$ tradeoff. In future, we plan to: (a) obtain tighter analytical lower bounds on the $ADD - P_{FA}$ tradeoff, and (b) design detection policies when the distributions $F_0(\cdot)$ and $F_i(\cdot)$ are unknown.

REFERENCES

- [1] I. F. Akyildiz, Won-Yeol Lee, M. C. Vuran, and S. Mohanty. A survey on spectrum management in cognitive radio networks. *IEEE Communications Magazine*, 46(4):40–48, April 2008.
- [2] Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995.
- [3] Blaise Kévin Guépié, Lionel Fillatre, and Igor Nikiforov. Detecting a suddenly arriving dynamic profile of finite duration. *IEEE Transactions on Information Theory*, 63(5):3039–3052, 2017.
- [4] L. Lai, Y. Fan, and H. V. Poor. Quickest detection in cognitive radio: A sequential change detection framework. In *IEEE Global Telecommunications Conference*, pages 1–5, Nov 2008.
- [5] Husheng Li, Chengzhi Li, and Huaiyu Dai. Quickest spectrum sensing in cognitive radio. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 203–208. IEEE, 2008.
- [6] James L. Massey. Optimum frame synchronization. *IEEE transactions on communications*, 20(2):115–119, 1972.
- [7] H Vincent Poor and Olympia Hadjiliadis. *Quickest detection*. Cambridge University Press Cambridge, 2009.
- [8] K Premkumar, A. Kumar, and V. V. Veeravalli. Bayesian quickest transient change detection. *Proceedings of Fifth International Workshop on Applied Probability (IWAP)*, 2010.
- [9] Matthijs T. J. Spaan and Nikos A. Vlassis. Perseus: Randomized Point-based Value Iteration for POMDPs. *CoRR*, abs/1109.2145, 2011.
- [10] Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. Chapman and Hall/CRC, 2014.
- [11] V. V. Veeravalli and T. Banerjee. Quickest Change Detection. *ArXiv e-prints*, October 2012.

Improved Data Fusion for Multi-Sensor Tracking using a Reinforced Viterbi Algorithm

Rajarshi Biswas, Akash S. Doshi, Akankshya Bhatta, Sibi Raj B. Pillai

Department of Electrical Engineering, Indian Institute of Technology Bombay

Email:{rajshree, akashdoshi, akanbhatta, bsraj}@ee.iitb.ac.in

Abstract—Employing multiple wide aperture radars with partially overlapping coverage to accurately track moving objects is becoming increasingly popular. However, identifying a common track across the radars can be challenging when each radar sensor obtains multiple measurements from different targets in its field of view. The presence of clutter and spurious measurements further complicates this problem. Data association and target tracking in this context can benefit from the combined processing of the sensor measurements. We adapt the well known single sensor Viterbi Data Association (VDA) algorithm to exchange information between multiple sensors, thereby reinforcing the target tracking performance. The proposed multi-sensor data fusion algorithm is demonstrated to have vastly improved performance over conventional single sensor techniques.

Keywords—Viterbi data association, multi-sensor, multi-track, data fusion, target tracking, radar signal processing.

I. INTRODUCTION

Identifying a common track from all participating radars in a multi-sensor multi-target environment is the main topic of this paper. Tracking of targets in clutter is very important for radar applications. In several situations, a single scan of the target scenario returns multiple measurements with uncertainty in the origin of such measurements. In this context, the data association (DA) procedure first ascertains whether a particular measurement can be associated with one of the target tracks, or is spurious in nature. DA for a single radar has been extensively studied in literature, popular algorithms include Probabilistic Data Association (PDA), Multiple Hypothesis Tracking (MHT), Fuzzy Data Association (FDA) as well as Viterbi Data Association (VDA) [1]. VDA has proved to be more effective at determining tracks at low Signal to Noise ratio (SNR) of the target measurements when false alarms due to clutter occur very often [2]. The VDA algorithm chooses a track by the optimization of a path cost metric over the entire observation window. It is a maximum likelihood approach to the data association problem [3], [4].

Tracking a single target using VDA (STVDA) was proposed in [5], where at most one measurement at any scan can belong to the target. When more than one target is in the view of the sensor, the STVDA will pick any one track based on the accumulated cost metric. In the presence of multiple-sensors observing the same target scene, each sensor, using VDA based tracking, may come up with a different optimal track. Identification of a common track among multiple sensors involves multi-sensor data association and data fusion.

Since multi-sensor data association is analogous to the track formation across consecutive scans of a sensor [6], it may seem that using the VDA on the assimilated measurements of participating sensors should prove advantageous. However, recall that the sensors may not always have the same measurement noise, and simply assimilating the measurements may end up favouring any one of the sensors. This reduces the robustness inherent to multi-sensor tracking in terms of observing a target scene [7]. Thus the tracking system should have a clever way of data association to effectively fuse the parallel measurements from the sensors.

There are well known methods for combining multi-sensor measurements from a single target, for example in [8]. Here, criticality depends on the underlying single track assumption. In particular, the fusion approach in [8] checks for the consistency of two measurements across the sensors based on the predicted tracks, before deciding to fuse the measurements. Note that each sensor makes only a single measurement. As opposed to this, a wide aperture radar may observe multiple targets in every scan at each sensor, thus data association followed by fusion becomes necessary for each track. This is a challenging problem. To the best of our knowledge data association and fusion for multiple targets using multiple sensors having partially overlapping fields of view have not been effectively addressed in literature in the VDA context. The key contribution of the current paper is a novel VDA where, DA at one sensor is reinforced using information exchanged from the other sensor, and vice-versa. Each sensor runs a separate VDA, but effectively makes use of the extrinsic information provided by the other sensor, reminiscent of the *turbo decoding* principle [9].

To incorporate extrinsic information exchange, an effective combination of the single target VDA [3], and multi-sensor data fusion technique of [8] is required. In our example systems, where target position and velocity are the two parameters of interest, the proposed algorithm accomplishes the identification of a common track based on both the kinematic parameters as well as the measurement amplitudes. The measurement amplitudes refer to the detected sample values obtained from the receiver, post matched-filtering. The amplitude values can be used to improve the reliability of tracking [10]. In a nutshell, a threshold based on the user specified probability of association is used to decide if a pair of measurements from the two sensors are to be associated to the same target. The individual tracks are extrinsically

reinforced as a function of the proximity of the estimates and the measurement amplitudes.

The organization of the paper is as follows. Section II describes the target scenario and system model. In Section III, we explain the existing VDA algorithm for a single sensor. We then present the reinforced VDA algorithm for multiple sensors in section IV. Simulation results are presented in Section V, demonstrating the utility of the proposed scheme. Section VI concludes the paper.

II. SYSTEM MODEL

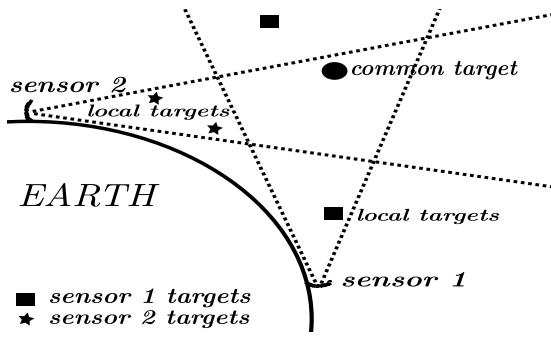


Fig. 1. Target scenario

Consider a scenario where a target is simultaneously observed by two sensors located in distinct geographic locations. Each radar performs M scans, and N measurements are acquired in each scan. It is assumed that one of the sensors, say *sensor 1*, has lower measurement noise compared to the other i.e. *sensor 2*. It is further assumed that the local target scene differs between the sensors, though there is partial overlap in the fields of view. It is assumed that a common track exists, measurements of which are being made by both the sensors. Notice the presence of additional tracks which are exclusive to a particular sensor. This is a common radar configuration for low altitude tracking, where the earth's curvature occludes each radar's field of view differently, see Fig 1. Each of the tracks is assumed to be from non-maneuvering point targets, with measurements corrupted by clutter or spurious measurements [11], a reasonable assumption for several practical target models.

A typical pair of multi-sensor measurements is illustrated in Fig 2, where each radar makes $M = 70$ scans. Each sensor observes $N = 5$ measurements at every sampling instant. Out of these, only one set of measurements correspond to a common track. Two of the remaining measurements at each sensor correspond to local tracks, and the remaining two correspond to clutter, which are chosen uniformly at random over the allowed range and velocity. For illustrative purpose, the common track is shown as blue in Fig 2, however this path is less obvious if the tracks are not highlighted. In fact, individual VDA at radar 1 may even fail to identify this common track as one of the possible tracks, this is shown in Fig 4a and 4b.

A. State Evolution and Multi-sensor Measurements

We describe a general model with S sensors employed for target tracking. Assume that the common track among the sensors evolves according to the state-space model,

$$\mathbf{x}_k = \Phi_k \mathbf{x}_{k-1} + \Gamma v_k, \quad (1)$$

where Φ_k is the state transition matrix, \mathbf{x}_k and \mathbf{x}_{k-1} are the target states at scans k and $k-1$ respectively. The state consists of the range and velocity of the target for a nearly constant velocity target model [12]. The process noise samples v_k are assumed to be white Gaussian with variance σ_a^2 representing acceleration noise. Γ is defined as $[\frac{1}{2}T_{mes}^2 \quad T_{mes}]^T$ where T_{mes} is the time interval between measurements [12].

Similarly, each local track in the exclusive field of view of sensor $q \in \{1, \dots, S\}$ is updated at scan k as,

$$\mathbf{x}_{k,j}^q = \Phi_k^q \mathbf{x}_{k-1,j}^q + \Gamma v_{k,j}^q, \quad (2)$$

where $j = 2, \dots, N_q$, and N_q is the number of targets in the view of sensor q . Note that except for the common track (taken as $j = 1$), $v_{k,j}^q, j \geq 2$ are independent across q , signifying the local tracks.

A sensor makes N measurements in each scan, i.e.

$$\mathbf{z}_{k,j}^q = \begin{cases} \mathbf{H}_k \mathbf{x}_{k,j}^q + \mathbf{w}_{k,j}^q, & \text{target originated} \\ \omega_{k,j}^q & \text{spurious,} \end{cases} \quad (3)$$

with $j = 1, 2, \dots, N$. Here $\mathbf{z}_{k,j}^q$ is the measurement, \mathbf{H}_k is the measurement matrix, and $\mathbf{w}_{k,j}^q$ is the measurement noise, assumed to be white Gaussian with covariance matrix \mathbf{R}_k^q .

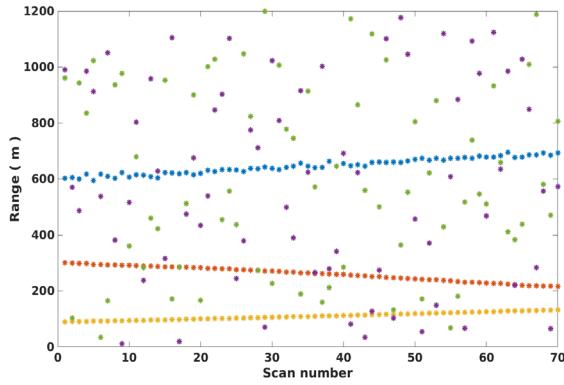
Remark. We have assumed suitable coordinate transformations to arrive at the linearised model as in (3). Geometric transformations have to be considered to accurately emulate Target scenario in Fig 1.

Each measurement $\mathbf{z}_{k,j}^q$ is also accompanied by a measurement amplitude, denoted as $\alpha(\mathbf{z}_{k,j}^q)$. Notice that these are the N detected measurements which were above the detection threshold. All targets are assumed to be **Swerling-2** in nature [12]. The amplitudes associated with measurements, both target originated and spurious, are as follows,

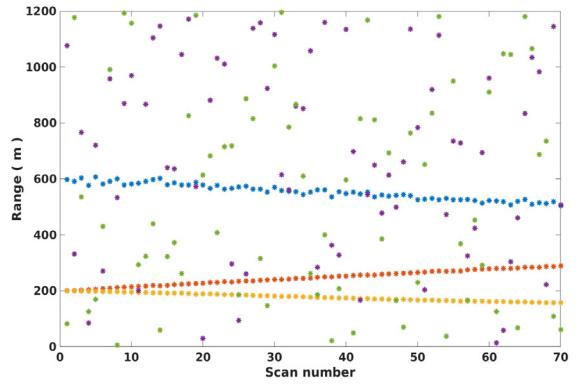
$$\alpha(\mathbf{z}_{k,j}^q) = \begin{cases} A_j^q + U_{k,j}^q & 1 \leq j \leq N_q, \\ \tilde{u}_{k,j}^q & N_q < j \leq N \end{cases} \quad (4)$$

where the amplitudes for the target originated measurements, i.e. $1 \leq j \leq N_q$, are perturbed about A_j^q randomly by $U_{k,j}^q$ a zero mean Gaussian random variable. Here $\alpha(\mathbf{z}_{k,j}^q)$ has a truncated (from below at zero) Gaussian pdf for $1 \leq j \leq N_q$ while the spurious amplitudes $\tilde{u}_{k,j}^q$ are drawn from independent Rayleigh distribution of scale parameter σ_j^q . Our choice of A_j^q , $U_{k,j}^q$ and σ_j^q are explained in more detail in Section V.

For the purpose of target-tracking and data fusion, the sensors exchange extrinsic information. In order to highlight the form and scope of this exchange, it is essential to understand the steps in the standard single sensor VDA, this is given in the next section.



(a) *Sensor 1*



(b) *Sensor 2*

Fig. 2. Target scenario measurements

III. VDA FOR A SINGLE SENSOR

In the VDA algorithm [3], [5], the target state estimates and the associated covariances are updated using a standard Kalman Filter [13], for each track of interest. These steps are performed sequentially for each scan, with each measurement associated to a track from the past.

Step 1: Initialisation

- The trellis has $N \times M$ nodes which are the measurements $\mathbf{z}_{k,j}$, shown in Fig 2 (range component only).
- The cost metric for each of the N tracks, and error covariances for the Kalman algorithm are initialized respectively to zero and identity.

Step 2: State transition metric and predecessor update

- For scan k , obtain the predicted state estimates from all updated state estimates at scan $k - 1$ i.e.

$$\hat{\mathbf{x}}_{k/k-1,i} = \Phi_k \hat{\mathbf{x}}_{k-1/k-1,i}, \quad 1 \leq i \leq N.$$

- At scan k , the residual covariances $\mathbf{S}_{k/k-1}$ are computed from nodes at scan $k - 1$, i.e.

$$\mathbf{S}_{k/k-1,i} = \mathbf{H}_k \mathbf{P}_{k/k-1,i} \mathbf{H}_k^T + \mathbf{R}_k,$$

where, $\mathbf{P}_{k/k-1,i} = \Phi_k \mathbf{P}_{k-1/k-1,i} \Phi_k^T + \mathbf{Q}_k$, is the error covariance matrix and \mathbf{Q}_k is the process noise covariance matrix.

- For $1 \leq j \leq N$, assuming a diffuse prior model [8], the transition metric $\bar{a}_{i,j}$ from node i to node j becomes [2],

$$\bar{a}_{i,j}(\mathbf{z}_{k,j}) \approx -\log \left(\frac{V_{i,j}}{N} P_d \mathcal{N}(\mathbf{H}_k \hat{\mathbf{x}}_{k/k-1,i}; \mathbf{S}_{k/k-1,i}) \right), \quad (5)$$

where $V_{i,j}$ is the volume of the validation region from the track ending in node i at scan $k - 1$, to node j at scan k . The probability of detection P_d is taken close to one. $\mathcal{N}(\mathbf{a}; \mathbf{B})$ stands for the normal distribution with mean vector \mathbf{a} and covariance matrix \mathbf{B} .

- The minimum cost path to j^{th} node from the previous scan instant is then indexed by,

$$\hat{i} = \underset{i}{\operatorname{argmin}} \bar{d}_{i,k-1} + \bar{a}_{i,j}. \quad (6)$$

- The cost or distance metric to j^{th} node is updated as $\bar{d}_{j,k} = \bar{d}_{\hat{i},k-1} + \bar{a}_{\hat{i},j}$ however, we modify the cost as,

$$\bar{d}_{j,k} = \bar{d}_{\hat{i},k-1} + \bar{a}_{\hat{i},j}. \quad (7)$$

Note that $\bar{a}_{\hat{i},j}$ here is a metric defined in Section IV (14).

- The minimizing index \hat{i} is stored for each node, to facilitate back-tracking from the last scan.

Step 3: Update Kalman estimator parameters

- For $1 \leq j \leq N$, update the Kalman gain \mathbf{K} using the apriori error covariance matrix of the predecessor node \hat{i} as follows.

$$\mathbf{K} = \mathbf{P}_{k/k-1,\hat{i}} \mathbf{H}_k^T \mathbf{S}_{k/k-1,\hat{i}}^{-1}, \quad (8)$$

$$\mathbf{P}_{k/k,j} = (\mathbf{I} - \mathbf{K} \mathbf{H}_k) \mathbf{P}_{k/k-1,\hat{i}}, \quad (9)$$

$$\hat{\mathbf{x}}_{k/k,j} = \hat{\mathbf{x}}_{k/k-1,\hat{i}} + \mathbf{K}(\mathbf{z}_{k,j} - \mathbf{H}_k \hat{\mathbf{x}}_{k/k-1,\hat{i}}). \quad (10)$$

Go back to Step 2 till all scans are considered.

After M scans, we can retrace one winning path from each measurement in the last scan, and arrange the N tracks in the ascending order of accumulated costs. Notice that we are only performing DA here.

IV. MULTI-SENSOR MULTI-TRACK FUSION

We now propose an effective algorithm for performing multi-sensor extrinsic information exchange to reinforce the track association at each sensor. The main idea is to modify the VDA cost metric to include extrinsic information exchange, this turns out to be very effective in recovering the common track.

To facilitate data fusion from two radars observing the same target, a threshold based track association was proposed in [8]. In this scheme, each radar obtains a single measurement in every scan, and a proximity metric is sequentially computed for the pair of measurements from the two sensors. If the

metric is below a threshold, the measurements are deemed to have a common origin, and these are fused to get a single estimate. While the essential characteristics of the proximity metric in [8] is retained in our algorithm here, we further adapt this metric to generate *extrinsic information* from each sensor, so as to reinforce the DA at the other sensor. In our scheme, each sensor individually runs the VDA as in Section III, except for using a modified cost metric in (6) – (7). Thus, we will only explain the computation of the reinforced cost metric. This involves three steps, (i) Inter-sensor DA, (ii) Extrinsic Reinforcement and (iii) Data Fusion. To keep the illustration straightforward, let us take two sensors, i.e. $S = 2$.

A. Inter-sensor DA

For each measurement $\mathbf{z}_{k,m}^q$, $m \in \{1, \dots, N\}$ at sensor q in scan k , recall that $\hat{\mathbf{x}}_{k/k,m}^q$ and $\mathbf{P}_{k/k,m}^q$ respectively denote the updated state estimate and error covariance matrix. The steps for inter-sensor DA are as follows [8],

- i) At scan k the difference of estimates is calculated as

$$\tilde{\mathbf{x}}_{m,n/k}^{1,2} = \hat{\mathbf{x}}_{k/k,m}^1 - \hat{\mathbf{x}}_{k/k,n}^2.$$

- ii) The covariance of the difference of estimates across sensor 1 and 2 for nodes m and n respectively is

$$\mathbf{P}_{m,n/k}^{1,2} = \mathbf{P}_{k/k,m}^1 + \mathbf{P}_{k/k,n}^2. \quad (11)$$

- iii) The squared Mahalanobis distance between the chosen estimates is computed as

$$D_{m,n/k}^{1,2} = (\tilde{\mathbf{x}}_{m,n/k}^{1,2})^T \{\mathbf{P}_{m,n/k}^{1,2}\}^{-1} \tilde{\mathbf{x}}_{m,n/k}^{1,2}.$$

- iv) Tracks association across sensors is now performed if, $D_{m,n/k}^{1,2} < D_{th}$, where $D_{th} = \chi_2^2(1 - \gamma)$, is the threshold below which the hypothesis that both $\hat{\mathbf{x}}_{k/k,m}^1$ and $\hat{\mathbf{x}}_{k/k,n}^2$ belong to the same target is true. γ is the user specified probability that the said hypothesis is false. $\chi_2^2(\cdot)$ is obtained from the Chi-square distribution table for two degrees of freedom [14], i.e. range and velocity.

B. Extrinsic Reinforcement

We will now explain the cost update procedure for the sensor $q = 1$, the other one is similar. Assign

$$m_j = \underset{m}{\operatorname{argmin}} D_{j,m}^{1,2} \text{ and } \hat{D}_j^{(q=1)} = D_{j,m_j}^{1,2}, \quad (12)$$

where ties are broken arbitrarily to choose a unique minimal argument. Thus, for node j in the first sensor, m_j is the proximal node from the second sensor. In order to incorporate measurement amplitudes, we first normalize the measurement amplitudes for sensor q as

$$\beta(\mathbf{z}_{k,j}^q) = \frac{\alpha(\mathbf{z}_{k,j}^q)}{\sum_{l=1}^N \alpha(\mathbf{z}_{k,l}^q)},$$

where $\alpha(\mathbf{z}_{k,j}^q)$ is the actual amplitude. On computing the transition metric in (5) for sensor $q = 1$, we first compute

$$b_{i,j}^q = \bar{a}_{i,j}(\mathbf{z}_{k,j}^q) - \underbrace{\log(\beta(\mathbf{z}_{k,j}^q))}_{\text{intrinsic reinforcement}} \quad (13)$$

An extrinsic reinforcement is now superposed on (13) to obtain a new effective transition metric

$$\bar{a}_{i,j}^q = \begin{cases} (b_{i,j}^q - \xi_j^q) & \text{if } \hat{D}_j^{(q=1)} < D_{th}, \\ b_{i,j}^q & \text{otherwise,} \end{cases} \quad (14)$$

where \hat{i} is that predecessor node in sensor q which leads to the least cost path to node j . The extrinsic function ξ_j^q is defined as,

$$\xi_j^q = \underbrace{\frac{\eta D_{th}}{D_{th} + \hat{D}_j^q} + \log \left(\frac{\max(\bar{\beta}, \beta(\mathbf{z}_{k,j}^q))}{\beta(\mathbf{z}_{k,j}^q)} \right)}_{\text{extrinsic reinforcement}}, \quad (15)$$

for $q = 1$, and $\bar{\beta}$ is the average amplitude, given by

$$\bar{\beta} = \frac{\beta(\mathbf{z}_{k,j}^1) + \beta(\mathbf{z}_{k,m_j}^2)}{2}.$$

In (15), η is a sensitivity parameter.

Once all the reinforced transition metrics $\bar{a}_{i,j}^q$ are obtained, the distance metric in equation (7) can be updated for node j at scan k as

$$\bar{d}_{j,k}^q = \bar{d}_{i,k-1}^q + \bar{a}_{i,j}^q. \quad (16)$$

While (13) ensures reduced transition cost thereby reinforcing a winning track (in VDA for a single sensor) based on measurement amplitudes alone, (14) and (15) ensure further reduction in transition cost as a function of proximity of state estimates i.e. \hat{D}_j^q provided IV-A(iv) is satisfied, when multiple sensors are involved. Multi-sensor amplitude information is also effectively incorporated in (15).

C. Data Fusion Step

The final step is to fuse the estimates from the proximal nodes j and m_j which qualify IV-A(iv) and equation (12), to obtain a better state estimate, as described in [8].

$$\begin{aligned} \hat{\mathbf{x}}_k^f &= \frac{\mathbf{P}_{k/k,m_j}^2 \hat{\mathbf{x}}_{k/k,j}^1 + \mathbf{P}_{k/k,j}^1 \hat{\mathbf{x}}_{k/k,m_j}^2}{\mathbf{P}_{j,m_j/k}^{1,2}}, \\ \mathbf{P}_k^f &= \frac{\mathbf{P}_{k/k,j}^1 \mathbf{P}_{k/k,m_j}^2}{\mathbf{P}_{j,m_j/k}^{1,2}}. \end{aligned}$$

We now set

$$\hat{\mathbf{x}}_{k/k,j}^1 = \hat{\mathbf{x}}_{k/k,m_j}^2 = \hat{\mathbf{x}}_k^f, \text{ and } \mathbf{P}_{k/k,j}^1 = \mathbf{P}_{k/k,m_j}^2 = \mathbf{P}_k^f.$$

These improved estimates $\hat{\mathbf{x}}_{k/k,j}^1$ and $\hat{\mathbf{x}}_{k/k,m_j}^2$ are used for state prediction in scan $k+1$. Thus, $\hat{\mathbf{x}}_{k+1/k,j}^1 = \hat{\mathbf{x}}_{k+1/k,m_j}^2$ while using the fused estimates for the next iteration.

V. RESULTS

A. Measurement generation

Simulations have been carried out for different scenarios, representative of the configuration shown in Fig 1. A typical set of measurements is shown in Fig 2. It is assumed that each scan returns $N = 5$ measurements. We take the state transition matrix Φ_k for the true tracks as

$$\Phi_k = \begin{bmatrix} 1 & T_{mes} \\ 0 & 1 \end{bmatrix}, \quad (17)$$

where T_{mes} is taken as 0.5 sec for our experiment. The measurement matrix \mathbf{H}_k is chosen as a 2×2 *Identity* matrix. The radar specifications considered for measurements generation is that of a short range X-band pulse Doppler radar with pulse width $\tau_{PW} = 0.2\text{ }\mu\text{sec}$, pulse repetition frequency of 2 kHz , dwell time $T_d = 16\text{ msec}$, with over-all system loss of 8.5 dB . The radar half-power beam-width is assumed 2.4° at 10 GHz with antenna gain of 37.45 dB . All targets have a radar cross section (RCS) of 0.001 m^2 .

Trajectory simulation parameters are given in table I.

TABLE I
TRAJECTORY PARAMETERS

q^{th} <i>sensor</i>	j^{th} track	σ_a^2	Init range (m)	Init vel (m/sec)
1,2	1	0.04	600 (w.r.t $q=1$)	2
1	2	0.02	300	-2
1	3	0.02	90	1
2	2	0.01	200	3
2	3	0.03	200	-1

The diagonal entries of the measurement noise covariance matrix \mathbf{R}_k^q are [15],

$$(\mathbf{R}_k)_{1,1}^{q=1,2} = \frac{c^2 \tau_{PW}^2}{8 \text{ SNR}}, \quad (\mathbf{R}_k)_{2,2}^{q=1,2} = \frac{(\frac{\lambda_t}{2})^2}{T_d^2(2 \text{ SNR})}. \quad (18)$$

where, λ_t is the transmission wavelength and c is the velocity of light. For the true tracks (indexed by measurements $j = 1, 2, 3$) in both the sensors, the mean amplitudes and the uniform perturbation range around it are generated as below.

The measurement amplitudes of spurious origin are drawn IID from Rayleigh distribution of scale parameter σ_j^q which is a function of the base-band noise power and the number of pulses integrated over T_d . The measurement amplitudes of target origin are perturbations about A_j^q , a function of the target RCS. The perturbations are modelled by $U_{k,j}^q$ a zero mean Gaussian random variable of variance which captures perturbations caused independently due to receiver noise and due to **Swerling-2** nature of the target. The measurement amplitudes are considered once the path loss has been accounted for by a range dependent gain [16].

TABLE II
PROCESS AND MEASUREMENT NOISE COVARIANCE MATRIX ELEMENTS

q	$(\mathbf{Q}_k)_{1,1}$	$(\mathbf{Q}_k)_{2,2}$	$(\mathbf{Q}_k)_{1,2} = (\mathbf{Q}_k)_{2,1}$	$(\mathbf{R}_k)_{1,1}$	$(\mathbf{R}_k)_{2,2}$
1,2	0.0006	0.01	0.0025	44.96	0.04391

B. Kalman Filter parameters for tracking

The Kalman Filter parameters, specifically the process noise covariance matrix for $\sigma_a^2 = 0.04$ and the measurement noise covariances at $\text{SNR} = 10\text{ dB}$ used for tracking at *sensor 1* and *sensor 2* are shown in table II.

C. Observation

We first demonstrate the results for a particular tracking experiment, which clearly highlights the advantage of the proposed scheme. The proposed scheme is tested on generated measurements with two sensors having the same specifications and the same tracking parametres. The threshold-based fusion for multi-sensor VDA correctly identified the common track (highlighted in blue) in Figure 5a and 5b. In the standalone mode of operation, with only intrinsic reinforcement at each sensor, the VDA preferred tracks with higher SNR (shown in blue) in Figure 4a and 4b. The extrinsic reinforcement provided in the current algorithm has allowed both the STVDA algorithms to converge to the same target.

In order to further check the robustness, each term of the measurement noise variance in (18) was identically scaled upwards for *sensor 2* simulating deterioration in SNR, and the common track identified for each case. Fig 3 shows the success rate, averaged over 20 independent trials.

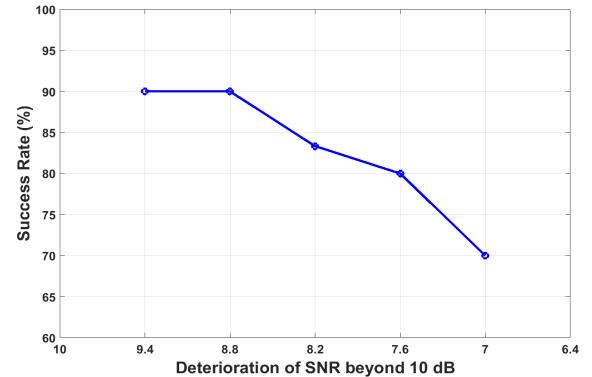


Fig. 3. Success Rate in Noisy Environments

D. Discussion

Conventional STVDA with intrinsic reinforcement alone ends up preferring target tracks with higher SNR often, losing out on the common track in this process, see Fig 4. The extrinsic reinforcement enables our algorithm to faithfully identify the common tracks. Notice that the algorithm requires specifying the sensitivity parameter η in (15). We have set $\eta = 30$ in our experiments based on empirical evidence, by progressively increasing η from *unity*, till a nominal common track is consistently identified.

A centralized data fusion architecture is assumed with the local trackers receiving feedback of the extrinsic information, the fused estimates and covariances, sequentially at every scan [8]. Once the common track is identified, subsequent local tracks can be targeted after subtracting the common one.

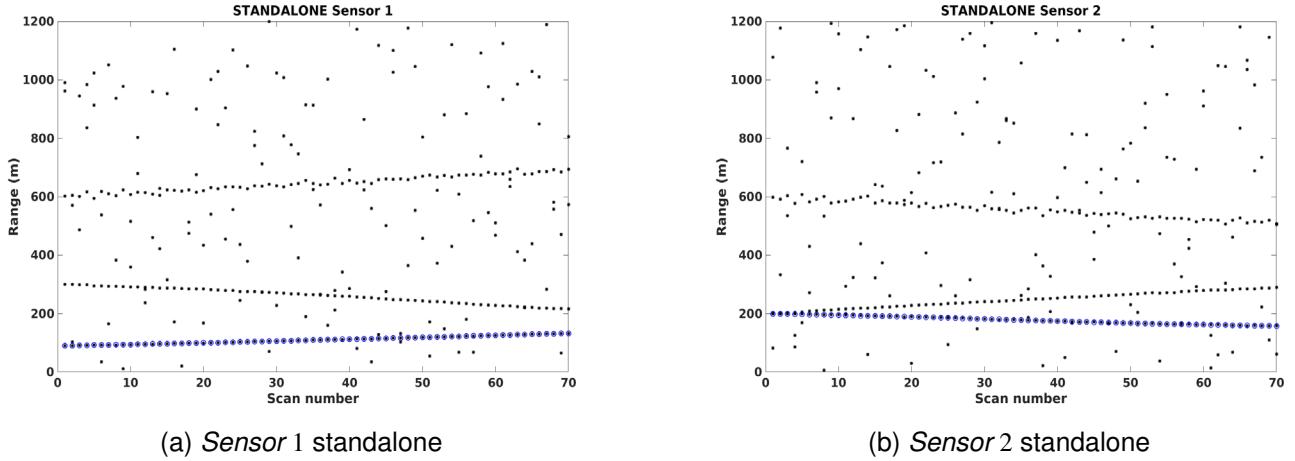


Fig. 4. STVDA with both sensors operating in standalone mode

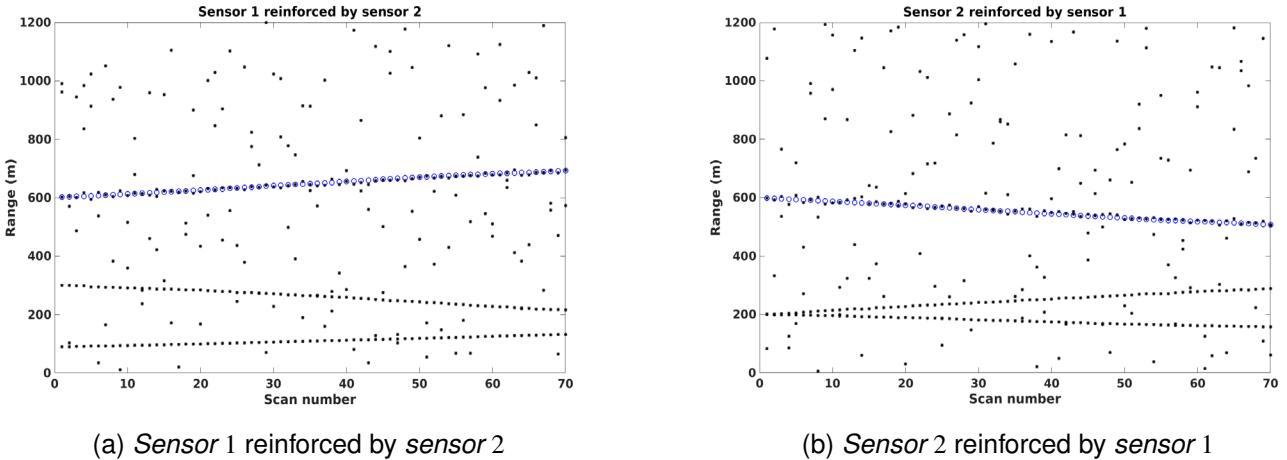


Fig. 5. Common track selection by Threshold-based Fusion for Multi-sensor VDA

VI. CONCLUSION AND FUTURE WORK

The algorithm proposed in the current paper is a promising approach to obtain common tracks among sensors in the VDA context. The algorithm uses all the features available from the target measurements typically available at a radar, and uses them to effectively converge to a single common track. Incorporation of the ability to track more than one common track is to be explored, in addition to a more structured choice for the sensitivity parameter η .

REFERENCES

- [1] A. Gad, F. Majdi, and M. Farooq, "A comparison of data association techniques for target tracking in clutter," in *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, vol. 2. IEEE, 2002, pp. 1126–1133.
- [2] B. La Scala and G. W. Pulford, "A viterbi algorithm for data association," in *Proc. International Radar Symposium*, vol. 3, 1996, pp. 1155–1164.
- [3] G. W. Pulford and B. F. La Scala, "Multihypothesis Viterbi Data Association: Algorithm Development and Assessment," *IEEE Transactions on Aerospace Electronic Systems*, vol. 46, pp. 583–609, Apr. 2010.
- [4] Y. Bar-Shalom, "Tracking methods in a multitarget environment," *IEEE Transactions on automatic control*, vol. 23, no. 4, pp. 618–626, 1978.
- [5] T. Quach and M. Farooq, "Maximum likelihood track formation with the viterbi algorithm," in *Decision and Control, 1994., Proceedings of the 33rd IEEE Conference on*, vol. 1. IEEE, 1994, pp. 271–276.
- [6] K. R. Pattipati, S. Deb, Y. Bar-Shalom, and R. B. Washburn, "A new relaxation algorithm and passive sensor data association," *IEEE Transactions on Automatic Control*, vol. 37, no. 2, pp. 198–213, 1992.
- [7] J. Llinas and D. L. Hall, "Multisensor data fusion," in *Handbook of multisensor data fusion*. CRC press, 2008, pp. 21–34.
- [8] Y. Bar-Shalom and X.-R. Li, *Multitarget-Multisensor Tracking: Principles and Techniques*. YBS London, UK:, 1995, vol. 19.
- [9] C. Berrou, R. Pyndiah, P. Adde, C. Douillard, and R. Le Bidan, "An overview of turbo codes and their applications," in *Wireless Technology, 2005. The European Conference on*. IEEE, 2005, pp. 1–9.
- [10] B. F. La Scala, "Viterbi data association tracking using amplitude information," in *Proc. 7th Int. Conf. Information Fusion*. Citeseer, 2004.
- [11] M. I. Skolnik, *Introduction to radar systems*. McGraw-Hill Education, 2002.
- [12] M. A. Richards, J. Scheer, W. A. Holm, and W. L. Melvin, *Principles of modern radar Volume I-Basic Principles*. Citeseer, 2010.
- [13] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory Algorithms and Software*. John Wiley & Sons, 2004.
- [14] G. J. Myatt, *Making sense of data: a practical guide to exploratory data analysis and data mining*. John Wiley & Sons, 2007.
- [15] S. Kingsley and S. Quegan, *Understanding radar systems*. SciTech Publishing, 1999, vol. 2.
- [16] M. A. Richards, *Fundamentals of radar signal processing*. McGraw-Hill Education, 2005.

Single versus Multi-Source Discrimination in Birdcalls using Zero-Frequency Filtering

Ragini Sinha, Vivek Vadluri, Ashish Arya, Padmanabhan Rajan

School of Computing and Electrical Engineering

Indian Institute of Technology

Mandi, India

Email: s16013@students.iitmandi.ac.in, padman@iitmandi.ac.in

Abstract—In the processing of bioacoustic recordings such as birdcalls, sometimes it is desirable to determine if a recording has one bird calling or has more than one. In this paper, we utilize the well-established zero-frequency filtering method, used for determining significant instants of excitation (also called epochs), for this task. By determining the average number of epochs per second, we are able to reliably discriminate birdcalls made by a single bird from those made by multiple birds. Experimental evaluation on three bioacoustic datasets confirms the reliability of the method. Species identification studies using deep neural network classifiers highlight the utility of the method.

Index Terms—Zero-frequency filter (ZFF); Hilbert envelope (HE); Bird call; Support vector machine (SVM); Deep neural network (DNN)

I. INTRODUCTION

Birds are useful indicators of environmental health. Their widespread roles in ecological functions include pollination, seed dispersion, and insectivory [1]. In recent years, many birds are on the edge of population decline due to human activities and climate change. Acoustic monitoring is a passive and effective way to monitor bird populations [1]. In general, acoustic monitoring of birds requires humans or algorithms to analyze recorded bird calls [3]. By this, several tasks can be performed, including species detection and classification, following migrant species and environmental health monitoring in regions of interest [1].

In many cases, the recording of the data is performed by automatic acoustic recorders. These devices can be programmed to record at preset time intervals. In field conditions, especially in high-biodiverse countries like India, many birds may be calling simultaneously. Unlike manually collected recordings, where highly directional microphones can be pointed to the bird of interest, automatic recorders collect all sounds in the environment. Further processing of such recordings, which may constitute a multitude of sources (i. e. many birds calling) or just one source (i. e. only one bird calling), has to be handled by appropriate techniques.

In this paper, we propose a procedure to determine if a bioacoustic recording has one source or more than one source. For instance, when applied to the identification of bird species from recordings collected automatically, such a system can be used to determine if the recording has to be processed by a single-label classifier, or a multi-label classifier. Another

application of such an algorithm is to search through a large archive of bioacoustic recordings.

The problem of determining the number of sources in a given signal has been investigated earlier, though in the more general context of audio source separation. Some of these methods are mentioned later in this section. In this paper, we adopt a signal processing approach, based on the fundamental method of sound production in birds.

The production of bird vocalization is similar to the production of speech in humans [8]. Airflow is established through the syrinx of the bird when it exhales. The syrinx is the vocal organ in birds. Unlike in humans, two pairs of labia are present between the bronchi and the vocal tract in birds. At certain conditions, oscillations are induced in the labia due to the airflow in the process of exhaling. This generates a pressure wave. When the pressure wave passes through the oroesophageal cavity and the trachea, the harmonic content of acoustic wave gets modified, which manifests in the sound signal produced [8]. The vocal tract plays an important role in the production of bird vocalizations. It impacts the relative magnitude of overtones [9]. Birds alter the vocal tract resonances to trace the fundamental frequency at the time of production of vocalization [9]. This leads to the similar functionality of the vocal tract in birds and humans.

The importance of the time instants of significant excitation in human speech production has been studied in various works [10], [15]. These instants of significant excitation have been used in various applications, including multispeaker separation [15] and foreground speech segmentation [17]. In this work, we utilize the instants of significant excitation in birdsong to determine if one or more than one bird is singing.

Several methods have been explored to detect the significant instants of excitation in human speech. In this work, we adopt the method described in [2]. The instants of significant excitation in the vocal tract are also termed epochs. The epoch location is useful for estimating the fundamental frequency very accurately. We extend this observation to determine the characteristics of the sound source with the knowledge of epochs [10].

Drawing on the method applied to human speech in [2], we assume that the resonances of the vocal tract produced during birdsong production extensively affect the location of the epochs. Therefore, we need a technique which removes the

effect of the resonances of the vocal tract. For this, we adopt the zero-frequency resonator described in the above paper. The zero-frequency resonator only allows the information centered around 0 Hz frequency and attenuates information at other frequencies. Hence, the resultant signal at the output of zero-frequency resonator has no effect of vocal tract resonances.

As in [2], we adopt the combination of the Hilbert envelope (HE) of the audio signal, followed by zero-frequency filtering (ZFF) for epoch detection from bird calls. The HE is used to strengthen the frequency component around the zero-frequency. The zero-frequency filter is used to detect the epoch locations from the HE. Once the epoch locations are determined, we compute the average number of epochs per second for the recording. We observe that the recordings having more than one bird calling occupy a higher range of the average number of epochs per second, in comparison to recordings having only one bird calling.

Related work: The problem of determining the number of sources can be seen as a part of the general source separation problem. In [14], a brief review of blind audio source separation is presented. In [12], a generalized mean shift clustering method was proposed for sound source separation in the time-frequency domain. The proposed idea was used for separating a number of sources from two mixtures. The idea performed well in both the cases i.e., linear speech mixture and binaural mixtures. In [13], a non-negative matrix factorization combined with a spectral mask was proposed for source separation. In the proposed method, speech and music signals were used to train the algorithm and a spectral mask was used to separate the single channel speech-music signal. In [15], a method using the information about the excitation source was proposed for separating speech of every speaker from multi-speaker speech, in a two-microphone case. Basically, the excitation source information was used to estimate the time delay between the two microphones. In the above works, the number of sources is determined automatically in the course of source separation.

Our study does not utilize multi-microphone audio. In our experiments, we make use of existing bioacoustic recordings collected by ecologists in field conditions.

The rest of the paper is organized as follows: Section II discusses the zero-frequency filter (ZFF) technique of [2] and the proposed framework. Section III discusses the experimental evaluations and section IV concludes our paper.

II. UTILIZING THE ZFF METHOD

ZFF method: This section is adapted from [2]. A zero-frequency filter is also termed as a zero-frequency resonator. The zero-frequency resonator has two poles located at 0 Hz. Both the poles are placed exactly at the unit circle in z-plane. Fig.1 shows the pole-zero plot and magnitude response of a zero-frequency resonator.

The transfer function for zero-frequency resonator is given as:

$$H_1(z) = \frac{1}{1 - 2z^{-1} + z^{-2}} \quad (1)$$

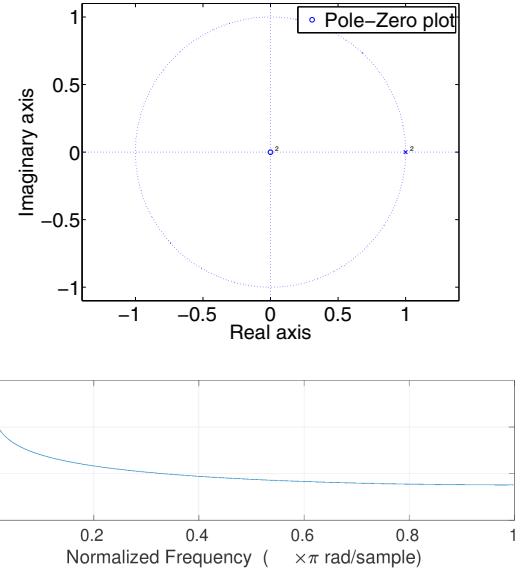


Fig. 1. The top image represents pole-zero plot and the bottom image represents the magnitude response of a zero-frequency resonator.

In the ZFF technique, first, the signal is differentiated and then is passed through the zero-frequency resonator two times. Due to sharp roll-off provided by the resonator, the output signal grows or decays with respect to time. For removing this trend from the resonator output, the mean over a short window is subtracted from each sample. The length of the window should be greater than one pitch period of the signal.

Fig.2 shows the block diagram representation of the proposed technique.

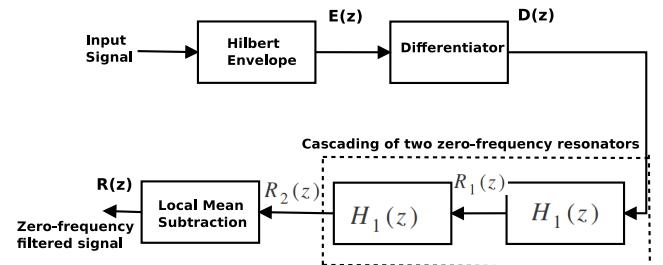


Fig. 2. Block diagram of the proposed technique. The HE of the input signal is passed through the ZFF to determine epoch locations.

Proposed framework: The proposed framework exploits the properties of the Hilbert envelope (HE) in the extraction of epochs in bird calls using the ZFF technique [2]. Data collected in field conditions include background noise, which may result in spurious epochs. To remove these spurious epochs, the zero-frequency filtering is performed on the HE of the signal.

In our experimentation, we have used a window length of approximately 1.2 times of one pitch period for de-trending the output of the ZFF. Zero crossings are computed on the de-trended signal [7]. Instants of significant excitation are present

at the positive zero crossings. To remove epochs with very low magnitude, a threshold of 0.01 is applied. The proposed method is shown in Fig. 2 and is illustrated in the time domain in Fig. 3.

The method is summarized below.

- The Hilbert envelope, $e(n)$ for bird call $c(n)$ is computed as:

$$e(n) = \sqrt{c^2(n) + c_h^2(n)} \quad (2)$$

Where $c_h(n)$ is the Hilbert transform of $c(n)$.

- The differenced signal $d(n)$ is computed as:

$$d(n) = e(n) - e(n - 1) \quad (3)$$

- $d(n)$ is passed through a cascade of two zero-frequency resonators.

$$r_1(n) = 2r_1(n - 1) - r_1(n - 2) + d(n) \quad (4)$$

$$r_2(n) = 2r_2(n - 1) - r_2(n - 2) + r_1(n) \quad (5)$$

- The local mean over a window length, approximately equal to 1.2 times of one pitch period is subtracted from each sample of $r_2(n)$.

$$r(n) = r_2(n) - \frac{1}{2P + 1} \sum_{k=-P}^P r_2(n+k) \quad (6)$$

Where, $2P + 1$ denotes the number of samples in one window length.

- To compute epochs locations, locations of positive zero crossings are computed.

After getting epoch locations, the audio recording is represented as a binary string, with 1 indicating epoch locations and 0 indicating no epochs. The epochs produced when a bird is calling does not vary rapidly. Hence the average number of epochs per second is relatively constant and is regularly spaced. However, when multiple birds call, the epochs in the resultant signal are irregularly spaced and have a higher value of average epochs per second. This is illustrated in Fig.4.

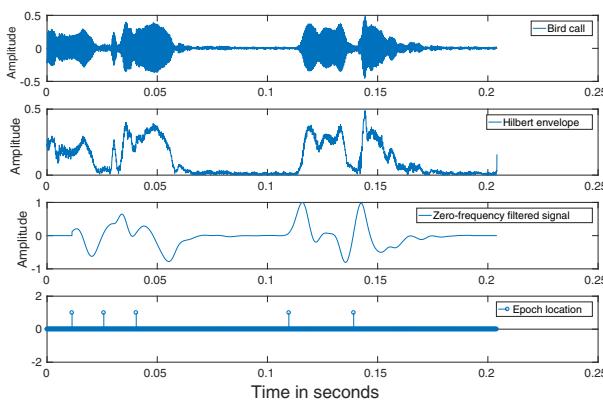


Fig. 3. From top to down, a waveform, its corresponding Hilbert envelope, resultant signal after zero-frequency filtering (ZFFS) and corresponding epoch locations are represented for a short segment of bird call.

III. EXPERIMENTAL EVALUATIONS

Experiments are conducted for three categories of bird calls i.e, single calls, partially mixed calls, and mixed calls. Recordings having only one bird calling is termed as single call recordings. Recordings having more than one bird calling but, with the call of one dominating the sound is termed as partially mixed call recordings (for example, when one bird is further away from the recorder.). The recordings having more than one bird calling, with all the sounds dominant are termed as mixed bird calls.

The experimental data are created using three datasets collected at various locations in a state in north India. Among these datasets, one dataset consists only single bird calls, one consists of partially mixed and mixed bird calls, the last one consists only mixed bird calls.

- GH dataset. This dataset consists of calls from 26 different species at a 44100Hz sampling rate, collected manually with a directional microphone. The GH dataset is clean and consists of calls of single birds, for 26 different species. We have a total of 586 wav files, considering all the species together. The single calls primarily come from this dataset. Partial calls and mixed calls are created from this dataset by artificially mixing two calls (sample wise addition in the time domain.)
- IM dataset consists of calls from 50 different species at a 44100Hz sampling rate, collected from a different location in the same state. Again, this was collected manually, but many recordings have several birds calling at the same time. This data contributed partially mixed and mixed calls.
Each audio file is manually trimmed to a length of 5 seconds.
- FD dataset consists of only mixed call recordings at various sampling rates. Some of these may have been collected by automated recorders, while some may have been collected manually (this information is not currently known.) Each wav file has different lengths. To use this dataset, we have re-sampled all the recordings at a 44100Hz sampling rate and trimmed all the recordings for 6 seconds.

The total number of wav files used for experiments for single bird calls, partially mixed bird calls and mixed bird calls are summarized in Table I.

Dataset	Wav files
Single calls	586
Partially mixed calls	367
Mixed calls	425

TABLE I
NUMBER OF TOTAL WAV FILES FOR SINGLE BIRD CALLS, PARTIALLY MIXED BIRD CALLS AND MIXED BIRD CALLS.

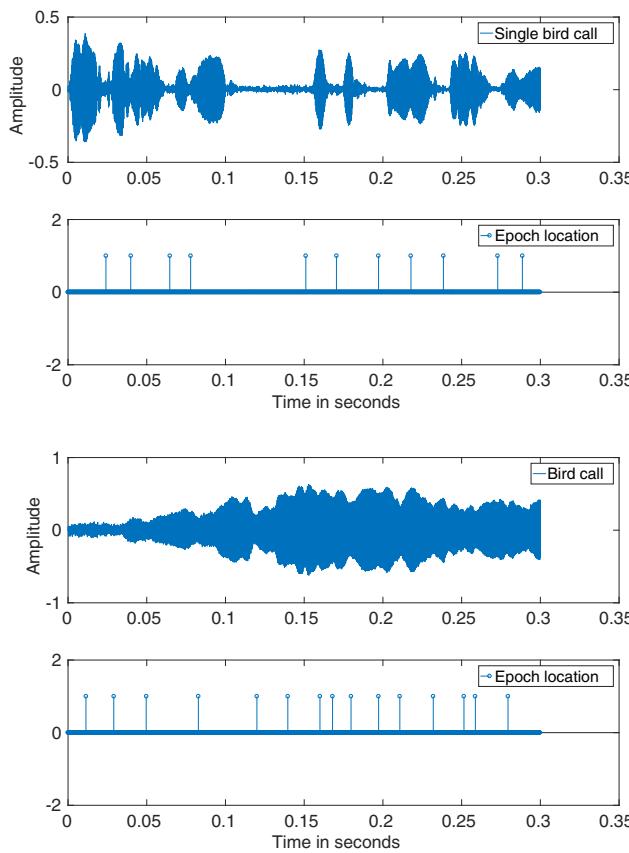


Fig. 4. Top figure represents the waveform and its corresponding epoch locations for single call and bottom figure represents the waveform and its corresponding epoch locations for mixed call.

A. Average number of epochs per second:

We have noted that the average number of epochs for mixed calls as well as for partially mixed calls are always greater than the number of epochs for single calls. This is true in the case of mixing the calls from the same species as well.

We have taken a short window length and computed the average value for that window length for the whole signal without any overlapping. The window length is taken as 1 second. We have computed the average for various windows like 5 milliseconds, 10 milliseconds, 50 milliseconds, 500 milliseconds and 1 second. For every size of the window length, we get the same range for the average number of epochs. But as window length has increased, we get more clarity in separating the ranges for one bird calling and more than one bird calling. Among all these window sizes, taking 1 second gives smoother level curve than other window lengths. This is illustrated in Fig.5. For this, we have taken 25 recordings from each type of call (single, partially mixed, and mixed.) Each recording is trimmed into a 6 seconds wav file. The computation of the average number of epochs per second is done and plotted to note the separation. The plots exhibit piece-wise nature because the average number of epochs is

constant per second and the windows do not overlap.

From Fig.5, it can be noted that the range for the average number of epochs per second for the mixed call is greater than the range for the average number of epochs per second for the single call and partially mixed call. It can also be observed that there is a very little amount of overlap between single calls and partially mixed calls. But, in the case of partially mixed calls and mixed calls, it can be noted that it is difficult to discriminate between the mixed call and partially mixed call due to the high amount of overlapping. Thus, the average number of epochs per second is a reliable method to identify single calls.

Fig.6 shows the relationship between the change in the average number of epochs per second with respect to the number of sources present in the recordings. The multi-source bird calls for two, three and four sources are created artificially using the single bird calls. We have considered 20 wav files for each type of multi-source bird calls to compute the average number of epochs per second. After computing the average number of epochs per second, the total sum of the average number of epochs per second for all the 20 wav file is computed for each case. The number of sources present in the recordings and the corresponding average number of epochs per second is shown in Fig.6. From the shown plot it is clear that the average number of epochs per second increases with an increasing number of sources.

B. Support vector machine (SVM) based discrimination of single-source and multi-source bird calls

We utilize a support vector machine (SVM) based classifier to classify the single-source bird calls and multi-source bird calls using the average number of epochs as features. For classification using the SVM, we have categorized our dataset in two classes, i.e., single-source having only single bird calls and multi-source bird calls, having partially and mixed both types of bird calls.

We have experimented the SVM based classification using both linear kernel and radial basis function (RBF) kernel. The SVM for both types of the kernel is trained on the binary representation obtained for epoch locations as input features. The input features are computed as the number of epochs per second having a shift of one sample. The SVM is trained on 70% of the data mentioned in Table I and tested on remaining data. In terms of the number of epochs per second with a shift of one sample, we have 3682 examples for training and 1579 examples for testing the SVM. For the linear kernel-based SVM we get the classification accuracy as 80.93% and for RBF kernel-based SVM we get 98.41% as the classification accuracy.

Hence, the SVM based discrimination is a reliable method to distinguish single calls from multiple calls.

C. Species identification studies

We utilize two deep neural network classifiers for species identification, each separately trained as a single-label classifier, and a multi-label classifier respectively. No attempt is

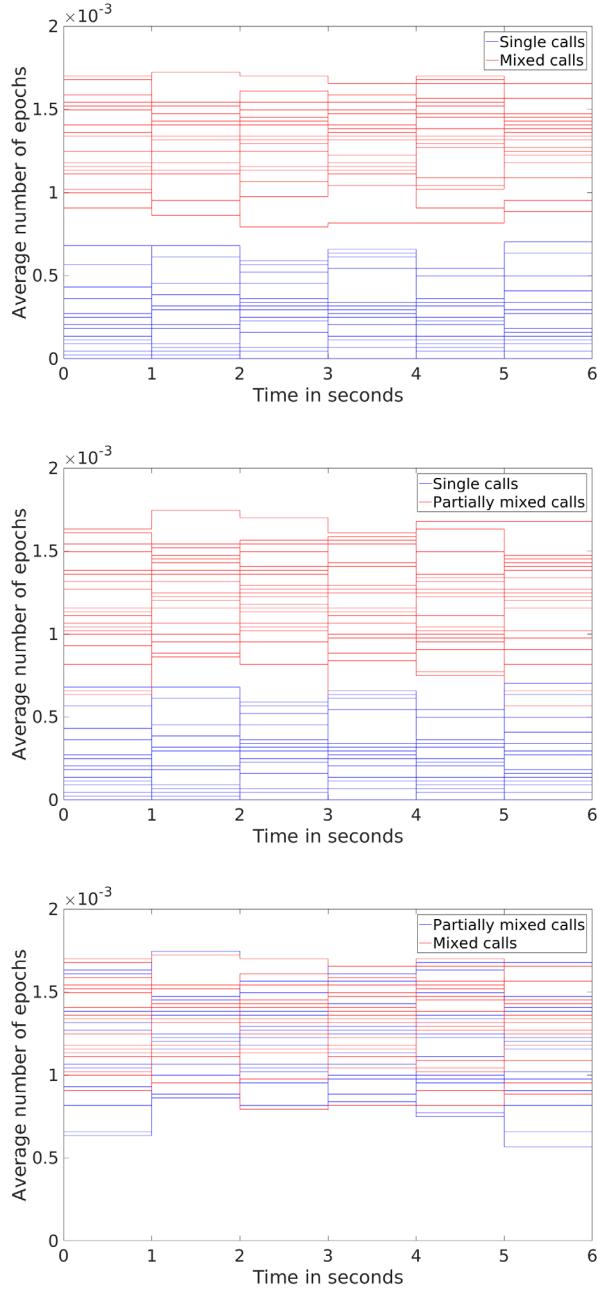


Fig. 5. Average number of epochs per second estimated from various types of calls. In each case, 25 calls are used. The top plot illustrates the clear separation between single calls and mixed calls. The middle plot shows the separation between single calls and partially mixed calls. The bottom plot shows that partially mixed and mixed calls are overlapping.

made here to compare the performance of these two DNNs. We simply wish to show the degradation of the performance of the single-label DNN when presented with multi-label data (i.e. recordings having multiple species calling at the same time.)

The single-label DNN is described in [11] is trained from the scratch while keeping all the hyper-parameters same as mentioned in the referred paper. The DNN is trained with 14

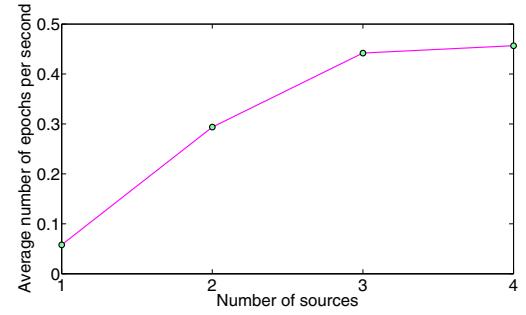


Fig. 6. Average number of epochs per second versus the number of sources available in the recordings. The X-axis represents the number of sources present in the recordings and Y-axis represents the average number of epochs per second. Each green dot in the plot shows the sum of the average number of epochs per second for all the 20 wav files for every case.

seconds of data per class from the GH dataset.

The amount of data used to train the multi-label DNN is far more non-uniform since the multi-class data is of partially mixed nature or mixed nature. The multi-label classifier uses data augmentation and multi-hot notations to train groups of bird species.

Both the single-label and the multi-label DNNs are used for 26 classes of species identification task. Both classifiers use Mel-frequency cepstral features (MFCC) (39 dimensional, with deltas and delta-deltas.) A context window of 7 frames is appended around every frame. MFCCs are computed with a frame size of 20 ms with a shift of 10 ms, using a Hamming window.

The results of classification are presented in Table II.

Classifier and data	Accuracy
Single-label DNN evaluated with single calls	98.0
Single-label DNN evaluated with mixed or partially mixed calls	26.1
Multi-label DNN evaluated with mixed or partially-mixed calls	81.4

TABLE II
CLASSIFICATION ACCURACY OF SINGLE-LABEL AND MULTI-LABEL DNN FOR SINGLE CALLS AND MIXED CALLS

Table II justifies the utility of the method described in this paper. Reliably distinguishing recordings with a single bird calling versus more than one bird calling is an important preprocessing step before employing the appropriate classifier.

IV. CONCLUSION

This work combines the utility of the Hilbert envelope and zero-frequency filtering to reliably determine if a bioacoustic recording has a single bird calling or has more than one bird calling. Utilizing the effectiveness of the ZFF technique, epoch locations are accurately determined from birdcalls. The average number of epochs per second reliably discriminates single calls from mixed or partially mixed calls. The limitation of the current method stems from that of the ZFF; it is not very robust to noisy conditions. Moreover, the method assumes

that only bird calls are present in the recording and no other sounds are present. But in general, the method illustrates the utility of determining the significant instants of excitation in bird calls, and how it can aid in their classification.

REFERENCES

- [1] T. Scott Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, no. S1, pp. S163–S173, 2008.
- [2] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.
- [3] I. Potamitis, S. Ntalampiras, O. Jahn and K. Riede, "Automatic bird sound detection in long real-field recordings: Applications and tools," *Applied Acoustics*, vol. 80, pp. 1–9, 2014.
- [4] T. Ananthapadmanabha, B. Yegnanarayana, "Epoch extraction of voiced speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no . 6, pp. 562–570, 1975.
- [5] Y. M. Cheng, D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no . 12, pp. 1805–1815, 1989.
- [6] C. Ma, Y. Kamp and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no . 2, pp. 285–265, 1994.
- [7] A. V. Oppenheim, "Discrete-time signal processing", 1999.
- [8] G. B. Mindlin, "The physics of birdsong production," *Contemporary physics*, vol. 54, no . 2, pp. 91–96, 2013.
- [9] S. Nowicki, "Vocal tract resonances in oscine bird sound production: evidence from birdsongs in a helium atmosphere," *Nature*, vol. 325, no . 6099, pp. 53, 1987.
- [10] B. Yegnanarayana, S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no . 5, pp. 651–697, 2011.
- [11] D. Chakraborty, P. Mukker, P. Rajan and A. D. Dileep, "Bird call identification using dynamic kernel-based support vector machines and deep neural networks," *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pp. 280–285, 2016.
- [12] D. Aylon, R. Gil-Pita, P. Jarabo-Amores, M. Rosa-Zurera, "Speech source separation using a generalized mean shift algorithm," *Signal Processing*, vol. 92, no . 9, pp. 2248–2252, 2012.
- [13] Grais, M. Emad, H. Erdogan, "Single channel speech-music separation using nonnegative matrix factorization and spectral masks," *Digital Signal Processing (DSP), 2011 17th International Conference on*, pp. 1–6, 2011.
- [14] E.Vincent, M. G. Jafari, S. A. Abdallah, M. Plumley and M. E. Davies, "Blind audio source separation," *Queen Mary, University of London, Tech Report C4DM-TR-05-01*, 2005.
- [15] B. Yegnanarayana, R. KumaraSwamy, and S. R. Mahadeva Prasanna, "Separation of multispeaker speech using excitation information," *ISCA Tutorial and Research Workshop (ITRW) on Non-Linear Speech Processing*, 2005.
- [16] S. Agnihotri, P. V. Sundeep, C. S. Seelamantula and R. Balakrishnan, "Quantifying vocal mimicry in the greater racket-tailed drongo: a comparison of automated methods and human assessment," *PloS one*, vol. 9, no . 3, 2014.
- [17] K. T. Deepak, B. D. Sarma, S. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

An Objective Measure to Assess Musical Noise using Connected Time-Frequency Regions

Ajey Saligrama

Dept of ECE,

PESIT-BSC,

B’lore, India

ajey@pes.edu

Ranjani H. G.

Dept of ECE,

Indian Institute of Science,

B’lore, India

ranjani@iisc.ac.in

R. Muralishankar

Dept of ECE,

CMRIT

B’lore, India

muralishankar@cmrit.ac.in

H. N. Shankar

Dept of ECE,

CMRIT

B’lore, India

hnshankar@cmrit.ac.in

Abstract—In this work, we propose an objective measure to assess the amount of musical noise that results from speech enhancement algorithms. The algorithms can result in non-smooth suppression of background noise which in turn translates to isolated regions of high energy, referred to as musical noise. We propose to identify such regions by combining time-frequency (TF) bins associated through connectivity along with additional properties of these regions such as area, aspect ratio and total energy. The objective measure proposed is based on density of such regions. The effectiveness of the proposed measure is studied by correlating it with subjective assessment of listeners using enhanced speech of various algorithms.

Index Terms—Musical Noise, Speech Enhancement, connected TF regions

I. INTRODUCTION

Enhancement of speech involves estimation of clean signal from its noisy counterpart. Enhancement algorithms make use of various models of noise and/or clean speech [1], [2], [3], [4]. Gain terms are estimated and used to suppress noise regions while retaining speech regions. If the estimated gain terms (corresponding to noise regions) are not smooth in either temporal or spectral domain, then they result in isolated regions of high energy in the time-frequency (TF) domain. These manifest as short bursts of tones to the listener. Further, these tones are positioned at various locations on the TF plane resulting in the annoying artifact referred to as ‘musical noise’ in the enhanced speech [5]. Figure 1(b) shows the spectrogram of enhanced speech obtained from iterative Wiener filter algorithm [6] - the dominance of musical noise can be visualized.

Many algorithms attempt to reduce musical noise by minimizing such isolated TF regions in the enhanced regions by introducing smooth gain factors along temporal [7] and/or spectral domain [8]. Figure 1(c) depicts a sample spectrogram using logMMSE algorithm [9] where the absence of isolated (high energy) TF regions can be observed from the spectrogram; hence, one can anticipate absence/reduction of musical noise (in the audio). This is at the cost of higher noise floor; an analysis and reasoning is detailed in [5].

Typical objective measures for assessing enhancement algorithms are segmental SNR (Signal-to-Noise Ratio) and PESQ (Perceptual Evaluation of Speech Quality) for quality attribute

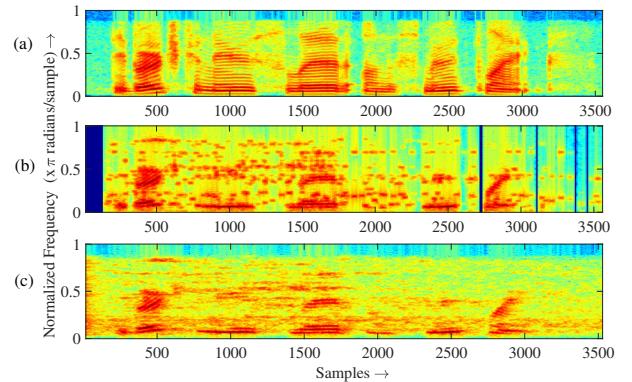


Fig. 1. Spectrogram of (a) clean speech utterance “The birch canoe slid on the smooth planks.” (b) enhanced speech using Iterative Wiener filter algorithm (c) enhanced speech using logMMSE algorithm. The input to both algorithms is 0dB noisy speech in car noise setting.

of test signal [10], and speech-based-STI (Speech Transmission Index), short-time objective intelligibility (STOI) measures for intelligibility attribute [11], [12]. Most of these measures are borrowed from speech coding and speech transmission domains, where one does not encounter musical noise artifacts; hence, these measures do not reflect degree of presence of musical noise.

While it is compelling to assess presence or absence of musical noise from visualization of the spectrogram, it has traditionally been assessed by subjective evaluations of the enhanced speech. We note that there have been a few attempts to assess musical noise in an objective way [13],[14],[15]. In [14], change in kurtosis is used as a measure to evaluate impact of isolated peaks while in [15] harmonicity of speech and similarity between successive frames are used as features to determine an objective measure. The authors of [13] quantify musical noise by detecting the number of isolated TF spots. The isolated TF spots are identified based on presence of high-energy areas through the zeros. Delaunay triangulation is used to connect all zeros and thus identify isolated regions of the spectrogram.

In this work, we propose an approach to identify regions corresponding to musical noise from the TF representation

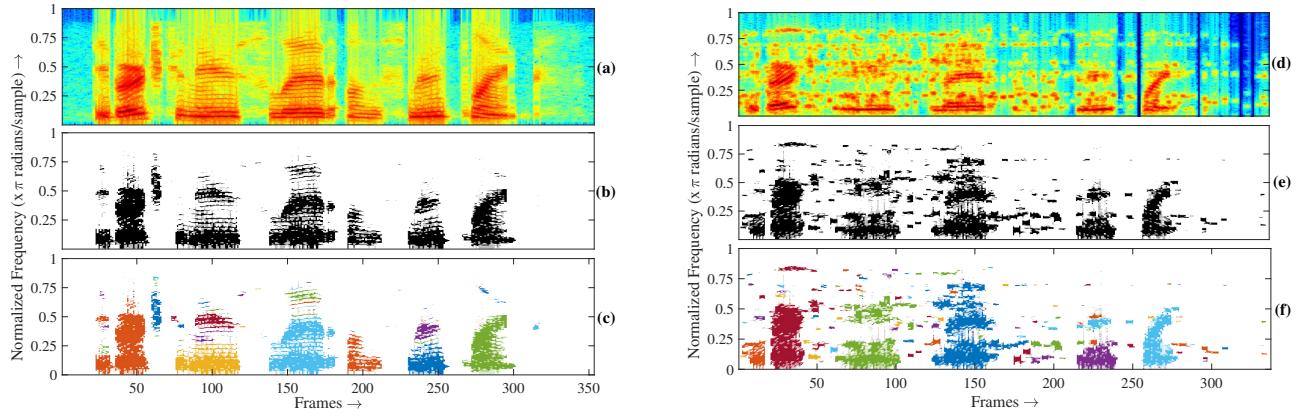


Fig. 2. (a) Spectrogram of an utterance of clean speech (b) Binary mask, B_c corresponding to clean speech (c) Component label matrix, C with different colors to depict various connected components, \mathcal{C}_l . (d , e, f) Spectrogram of \hat{S} from Iterative Wiener filter algorithm with noisy speech in 0dB car noise and its corresponding binary mask, B and component label matrix, C respectively.

of enhanced speech. TF bins (of enhanced speech) that can be associated through connectivity are combined to form candidate musical noise (TF) regions. Some attributes of these TF regions are used to determine the correspondence to speech/musical noise. An objective measure is proposed based on the collection of such TF regions. This measure in turn is used to quantify musical noise present in speech enhancement algorithms such as spectral subtraction [16], iterative Wiener filter [6], logMMSE [9] and Spectro-Temporal Discriminative Random Fields (ST-DRF) [17]. The assessment is correlated with subjective assessment of listeners to understand its effectiveness as an objective measure.

The work is organized as follows: Section II outlines our proposed approach. Section III evaluates the proposed objective measure on the enhanced speech obtained from enhancement algorithms, followed by the correlation analysis between subjective and objective assessments. We further explore the possibility to suppress these musical noise regions using a CASA (Computational Auditory Scene Analysis)-like binary gain approach (Section IV).

II. CONNECTED TF REGIONS FOR MUSICAL NOISE DETECTION

Let $\hat{s}(n)$ be the enhanced speech from any enhancement algorithm. Let its equivalent (and invertible) TF representation be $\hat{S} = \hat{S}(n, k), \forall n \in [1, N], k \in [1, K]$ where N and K are the number of temporal frames and spectral components of the representation respectively. Let $B(n, k)$ be a binary mask associated with the spectro-temporal bin, $\hat{S}(n, k)$, such that:

$$B(n, k) = \begin{cases} 1 & \text{when } \hat{S}(n, k) \text{ has high energy,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Essentially, $B(n, k)$ is a binary mask to be estimated so that the TF components of test signal can be separated into foreground and background components. We assume the first few frames of the signal correspond to background noise and

hence, we propose to infer B from the observed \hat{S} (and similarly binary mask, B_c from clean speech) as:

$$B(n, k) = \begin{cases} 0 & \text{when } \hat{S}(n, k) \leq I_{thr}, \\ 1 & \text{otherwise } \forall n, k. \end{cases} \quad (2)$$

where I_{thr} is the average energy of a TF bin in the first few frames of $\hat{S}(n, k)$. This gives us a binary TF representation of high energy regions. In order to identify isolated regions from this, we propose to find and label the connected TF bins of B similar to connected component labeling of a binary image [18]. This requires labeling of connected foreground TF bins. We use the flood-fill algorithm which is detailed below.

A. Flood-fill algorithm for labeling

Let C be the component label matrix corresponding to the unlabeled foreground regions B . We obtain connected component label in an iterative procedure using the flood fill algorithm. In this algorithm, the first encountered foreground TF bin is assigned a label. The foreground neighbors of labeled foreground TF bins are also assigned the same label. This label assignment continues until there are no more foreground neighbors for the said connected component. Any foreground TF bin that is parsed once is set as background to ensure no TF bin is parsed twice. This is repeated across all the foreground TF bins. The algorithm is detailed in Algorithm 1.

B. Features from connected components

Let \mathcal{C}_l be set of all TF bins corresponding to connected component label l and C represent $\{\mathcal{C}_l\}, \forall l$. Thus, C , contains all (component labeled) foreground TF regions of \hat{S} and is well above a threshold. An example of connected component labeling is shown in Figure 2 for clean speech data and enhanced speech from iterative Wiener Filter, \hat{S} as input.

Let $M \subset C$ be set of the connected components corresponding to musical noise. In order to estimate M from C , we

Algorithm 1 Flood fill based connected component analysis

Require: $C(n, k) = 0, \forall n, k$
Require: Init $l = 1$
for $n = 1$ to N **do**
for $k = 1$ to K **do**
 $P \leftarrow \{(n, k)\}$
if $B(P)$ is foreground TF bin **then**
while $|P| > 0$ **do**
 $C(P) \leftarrow l,$
 $B(P) \leftarrow 0,$
 $P \leftarrow \{ \text{Foreground Neighbors}(P) \}$
end while
 $l \leftarrow l + 1$
end if
end for
end for

propose to utilize the following attributes of each component, $C_l, \forall l$ of \mathbf{C} as features:

- Area of connected components in terms of TF bins
- Total energy of each of the connected region
- The ratio of TF bins along temporal (X) to that along spectral (Y) regions

These attributes are based on observations wherein the connected components are more in number and of smaller surface area in random regions of TF plane when there is presence of musical noise. Although, these attributes are simple and intuitive, we bear in mind that it is not straight-forward to use these in a training-test based classification setup. This is because these connected components of \mathbf{C} obtained from $\hat{\mathbf{S}}$ are not labeled as speech or musical noise. We design a way to obtain labels from such a scenario and detail the same ahead.

C. Identifying speech/musical noise regions

In order to label the connected components of \mathbf{C} , we use clean speech for the training phase. We propose to mask out TF bins corresponding to those temporal frames of \mathbf{B} for which \mathbf{B}_c has a foreground component. If we represent this new mask as \mathbf{W} , then we have:

$$W(n) = \begin{cases} 0 & \text{if any } B_c(n, k) = 1, \forall k \\ B(n) & \text{otherwise } \forall k \end{cases} \quad (3)$$

with $W(n)$ (or $B(n)$) denotes (all) the spectral components of n^{th} frame. This results in a clean speech guided VAD (voice activity detector) for $\hat{\mathbf{S}}$. Figure 3 shows one such sample of \mathbf{W} which corresponds to a portion of musical noise in $\hat{\mathbf{S}}$. Thus, we can, for all practical purposes, consider connected components from \mathbf{W} as corresponding to musical noise and those obtained from \mathbf{B}_c as that of speech. We extract the attributes of connected components and use these as features. A scatter plot of these attributes for connected components corresponding to speech and musical noise labels (obtained

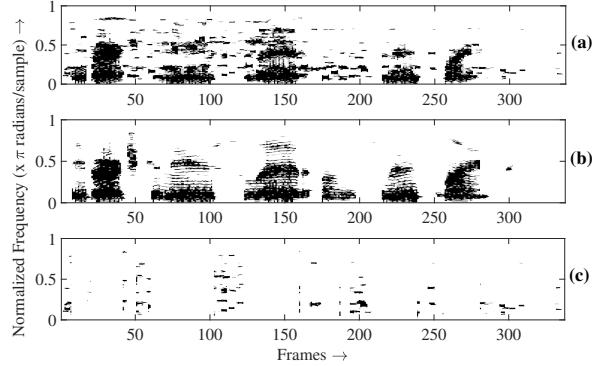


Fig. 3. Binary masks corresponding to (a) $\hat{\mathbf{S}}$ (b) \mathbf{S} (c) \mathbf{W} estimated from clean speech guided VAD on $\hat{\mathbf{S}}$.

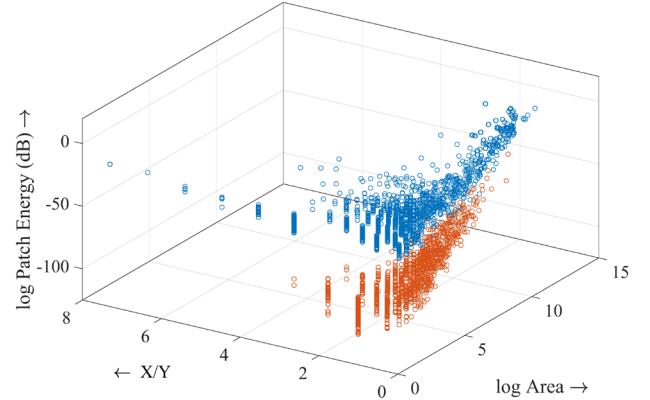


Fig. 4. Scatter plot of feature vectors of clean speech (in blue) and musical noise (in red).

from iterative Wiener filter algorithm for speech corrupted by 0 dB car noise) is shown in Figure 4.

1) *Training test set-up:* We train the model for speech corrupted by 0 dB car noise. The labels are obtained using $\hat{\mathbf{S}}$ from iterative Wiener filter. A 5-fold training-test setup is considered; the training data for clean speech is the connected components from \mathbf{S} while that for musical noise are the connected components from estimated \mathbf{W} . A linear discriminator is used to model the two classes. For testing, we consider the connected components of $\hat{\mathbf{S}}$ and classify them to speech/musical noise. Thus, \mathbf{M} is estimated for each utterance as the set of connected components corresponding to musical noise.

For testing performance on other noise conditions and algorithms, we fix the model obtained from iterative Wiener filter for speech corrupted by 0 dB car noise. The connected components of $\hat{\mathbf{S}}$ are classified as speech/musical noise.

We recognize the non-availability of clean speech for obtaining the mask \mathbf{W} , in a practical speech enhancement scenario and hence advocate fixing the model obtained from data set which have availability of clean speech.

2) *Quantify musical noise:* In order to quantify the amount of musical noise in an enhanced speech signal, we propose

the following estimators:

- The density of estimated TF connected regions corresponding to M per second as the objective measure i.e., $N = |M|/T$ where T is duration of the utterance. In order to quantify the amount of musical noise arising from an algorithm, we use mean and standard deviation of N across the utterances of the considered dataset as a representative of the measures; we use N_{avg} to depict the same.
- Percent of connected components classified as musical noise i.e., $|M|/|C|$.
- Total confidence measure of all elements of M - obtained by summing all the probabilities of components classified as musical noise.

III. EVALUATION OF OBJECTIVE MEASURE USING CONNECTED TF REGIONS

In this section, we evaluate the proposed objective measure, N on the NOIZEUS dataset. The dataset consists of 30 sentences from 6 speakers. The sampling frequency is 8 kHz. We consider 4 noise conditions: AWGN (Additive White Gaussian Noise), car, street and train noise. The noisy speech is at 0 dB, 5 dB, 10 dB and 15 dB signal-to-noise-ratio (SNR). For the TF representation, the enhanced speech is (Hamming) windowed and buffered with 32 ms of data and 75% overlap and 512 point FFT.

We first conduct listening tests to study correlation between proposed measures and subjective ratings.

A. Test on clean speech

The features are proposed to identify TF regions containing musical noise; however, we also check the baseline performance using clean speech as test input i.e., we estimate N with S as input. We expect the features to be robust so that the objective measure shows low values of N when the input is clean speech. However, it is not possible that $N = 0$ as there exists high frequency components corresponding to short duration regions; it is possible that these could be categorized as musical noise.

B. Correlate with subjective assessments

Listening tests are conducted for subjective evaluation of musical noise present in \hat{S} estimated from various algorithms. Ten subjects with normal hearing are considered for the listening tests. JBL-T250SI headphones are used. Each subject is presented with 8 experiments; each experiment involves comparing enhanced speech from 4 algorithms - such as spectral subtraction [16], iterative Wiener filter [6], logMMSE [9] and ST-DRF [17] each having input speech corrupted by AWGN at two SNR levels (0 and 5 dB). Each noise condition is evaluated by a listener with 4 different utterances - 2 utterances each by male and female speakers.

The subjects go through a musical noise familiarization phase; audio files with only musical noise, clean speech, enhanced speech are played, thus familiarizing with different degrees of musical noise. This is followed by a MUSHRA

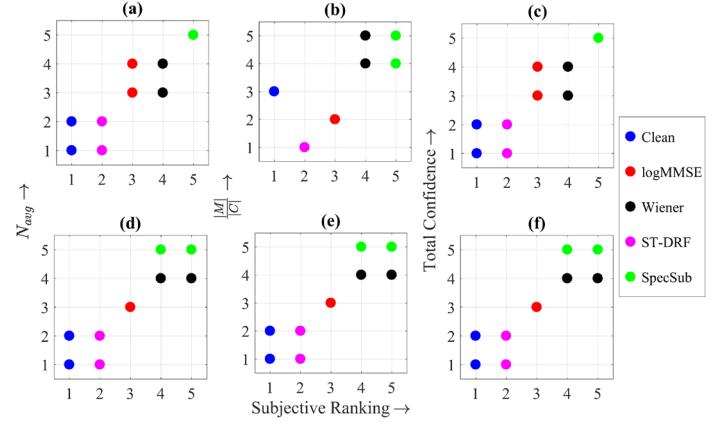


Fig. 5. Correlation between rankings obtained from subjective listening tests and proposed objective measures. 4 speech enhanced files from 4 different algorithms (and clean speech) are ranked for presence of musical noise (rank 1 implies least musical noise is detected). Sub-plots (a,b,c) correspond to 0dB AWGN noise setting as input to algorithms while (d,e,f) correspond to 5dB AWGN noise setting.

evaluation [19]. The subjects are asked to listen to the amount of musical noise, compare and rate the outputs of the 4 algorithms along with a hidden reference in each experiment. The ratings are on a scale of 0-100 with 100 representing no musical noise, with the clean speech provided as a reference for absence of musical noise. Only those scores which have correctly identified the hidden reference with a score of 100 are used to screen listeners' ratings. The scores are averaged for each utterance across the listeners.

We then convert the subjective ratings to subjective rankings of the algorithms. This in turn is correlated with the rankings (of the algorithms) that can be inferred from the objective measures estimated from each utterance. The ranking ranges from (1-5) where 1 implies algorithm with minimum musical noise and 5 implies that with maximum musical noise. Subjective and objective rankings for 2 different input SNR in AWGN setting are shown in Figure 5. Maximum correlation implies all dots must lie on or close to the diagonal.

All listeners have always correctly identified clean speech in subjective listening tests and thus clean speech is ranked for minimum musical noise; whereas, the objective measure shows that clean speech can sometimes be ranked lower than enhanced speech from ST-DRF. This can be attributed to lower presence of isolated components in the ST-DRF enhanced signal owing to spectro-temporal smoothing in the formulation.

Amongst the objective measures considered in Section II-C2, we observe that N_{avg} and total confidence measure show similar correlations to subjective rankings. We also observe that N_{avg} is better correlated to subjective ratings than $\frac{|M|}{|C|}$ - this deviance could be due to increased presence of isolated harmonic patches in clean speech; a normalization with respect to $|C|$ can result in higher ratio than that obtained from enhanced speech.

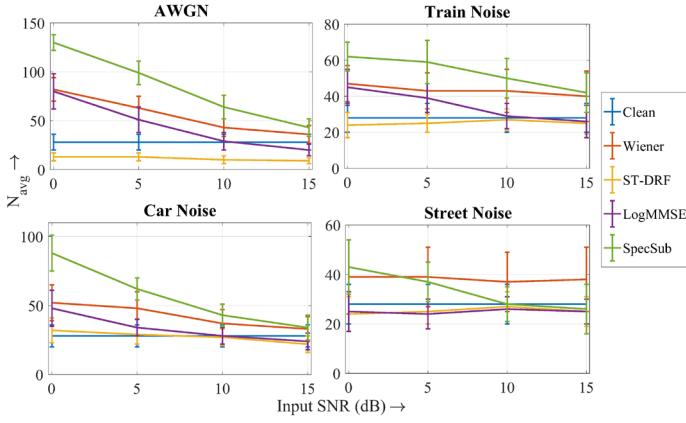


Fig. 6. Average density (and standard deviation) of musical noise components, N_{avg} measured by the proposed approach with enhanced speech from various algorithms in various noise settings.

C. Performance in various noise conditions

In this section, we measure the amount of musical noise using N_{avg} ; enhanced speech obtained from various enhancement algorithms are used as the test input. Again, we consider algorithms - Spectral subtraction, iterative-WF, logMMSE and ST-DRF. Figure 6 shows N_{avg} of connected TF regions of enhanced speech classified as musical noise in various noise conditions. We observe that ST-DRF algorithm results in minimum musical noise artifact in all noise conditions which can be attributed to the tempo-spectral continuity incorporated in the DRF framework. N_{avg} from spectral subtraction is maximum under all noise conditions. Between iterative Wiener filter and logMMSE algorithms, the latter has lesser musical noise in all noise settings. One of the reasons for better performance of logMMSE algorithm over iterative Wiener filter is that the gain factors introduced has a minimum noise floor whereas in iterative Wiener filter, TF regions corresponding to noise undergo maximum attenuation [5]. As a result, uneven gain factors are more pronounced with iterative Wiener algorithms than in logMMSE algorithm.

However, the presence of the noise floor in logMMSE algorithm will have an effect in detecting connected components in the enhanced signal. Though the noise floor of logMMSE will mask residual artifacts/musical noise, the same is not necessarily reflected in estimated \mathbf{B} .

IV. SUPPRESSING MUSICAL NOISE

The proposed algorithm detects the musical noise TF regions of enhanced speech. These detected regions can in-turn be suppressed. Thus, the algorithm can be used as a post-filter for enhancement algorithms. We use \mathbf{M} to create a binary mask to suppress these regions i.e.,

$$\hat{S}_{pf}(n, k) = BM(n, k) * \hat{S}(n, k) \quad (4)$$

where,

$$BM(n, k) = \begin{cases} 0 & (n, k) \in \mathbf{M}, \\ 1 & \text{otherwise.} \end{cases} \quad (5)$$

Figure 8 shows quality and intelligibility performance before and after the proposed post-processing. The input to various algorithms is noisy speech at 0 dB in various noise settings. We observe that there is not significant change in quality between \hat{S} and \hat{S}_{pf} ; the intelligibility scores shows dip across all algorithms. Listeners have reported reduced musical noise. Sample audio files have been uploaded (<https://tinyurl.com/y9kt4axp>) for sample informal listening. Figure 7 shows a sample spectrogram of the post processed enhanced speech. We observe that the musical noise regions are reduced.

We have also considered clean speech as input and observed that suppressing regions classified as musical noise results in a dip PESQ and STOI scores to 3.3 and 0.6 (from 4.5 and 1 respectively); however, we have observed through informal listening that there is no loss in intelligibility though slight quality degradation can be perceived.

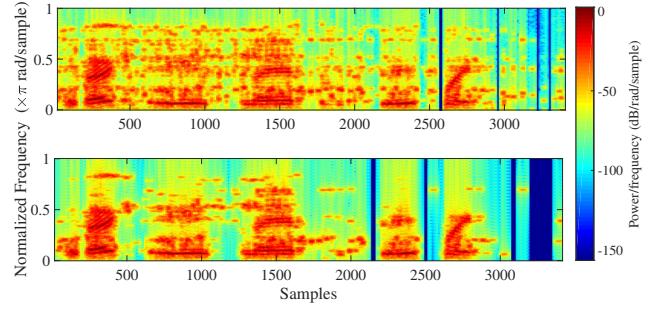


Fig. 7. Spectrogram of (a) enhanced signal from iterative Wiener filter algorithm (with noisy speech in car noise at 0dB) (b) post processed speech with above enhanced speech as input resulting in musical noise suppression.

V. CONCLUSIONS

We have addressed the challenge of evaluating musical noise in speech enhancement algorithms. We have used connected TF regions of the spectrogram to identify possible region candidates. Attributes of the region, such as the area in terms of TF bins, ratio of spread of connected TF regions along temporal to that along spectral dimension, the energy of the TF regions, are used to estimate musical noise regions. The density of such regions is proposed as an objective measure. We correlate the rankings obtained from the proposed objective measure with that of subjective measures. Correlation is found to be higher for lower input SNR cases. In addition, as a post-processing step to speech enhancement algorithms, we use these estimated musical noise TF regions to suppress the artifact itself.

REFERENCES

- [1] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

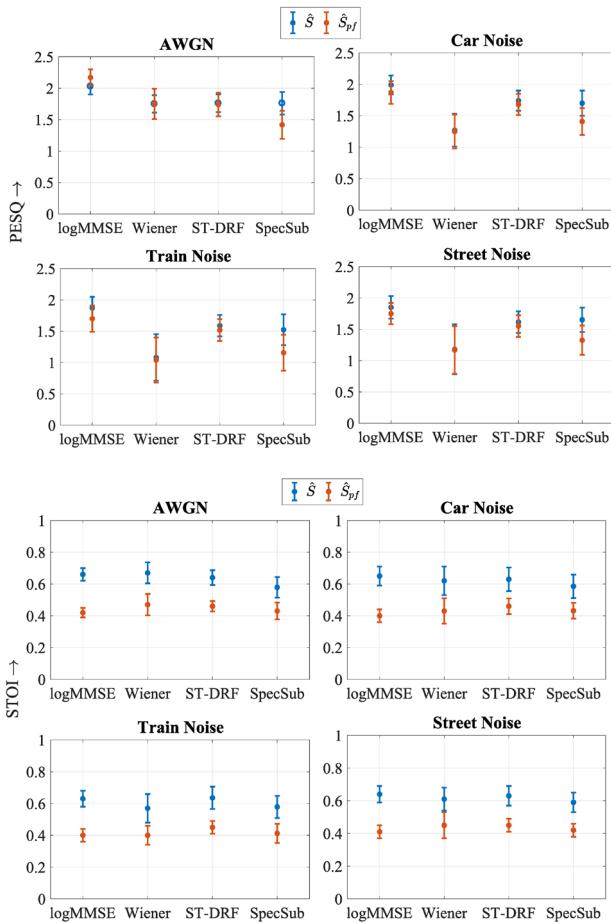


Fig. 8. Quality and intelligibility of \hat{S} (blue) and \hat{S}_{pf} (red) of various algorithms using PESQ and STOI measures.

evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. IEEE, 2001, pp. 749–752.

- [11] R. L. Goldsworthy and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America (JASA)*, vol. 116, no. 6, pp. 3679–3689, 2004.
- [12] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Intl. Conf. on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4214–4217.
- [13] R. Hamon, V. Emiya, L. Rencker, W. Wang, and M. Plumbley, "Assessment of musical noise using localization of isolated peaks in time-frequency domain," in *Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 696–700.
- [14] T. Inoue, H. Saruwatari, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in wiener filter via higher-order statistics," 2010.
- [15] N. Derakhshan, M. Rahmani, A. Akbari, and A. Ayatollahi, "An objective measure for the musical noise assessment in noise reduction systems," in *Acoustics, Speech and Signal Processing, IEEE Int'l. Conf. on (ICASSP)*. IEEE, 2009, pp. 4429–4432.
- [16] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. IEEE, 1979, pp. 208–211.
- [17] A. Saligrama, H. G. Ranjani, H. N. Shankar, and R. Muralishankar, "Speech enhancement using discriminative random fields," in *Proc. TENCON, IEEE Region 10 Conf.* IEEE, 2015, pp. 1–6.
- [18] L. G. Shapiro, "Connected component labeling and adjacency graph construction," in *Topological Algorithms for Digital Image Processing*, ser. Machine Intelligence and Pattern Recognition, T. Y. Kong and A. Rosenfeld, Eds. North-Holland, 1996, vol. 19, pp. 1 – 30.
- [19] I. Recommendation, "BS. 1534-1. method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union, Geneva*, 2001.

- [2] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–253.
- [3] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *Speech and Audio Process., IEEE Trans. on*, vol. 13, no. 5, pp. 870–881, 2005.
- [4] N. R. Muraka and C. S. Seelamantula, "A risk-estimation-based comparison of mean square error and Itakura-Saito distortion measures for speech enhancement," in *Proc. of INTERSPEECH, Italy*, 2011, pp. 349–352.
- [5] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 2007.
- [6] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, 1978.
- [7] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE Int'l. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4. IEEE, May 2002, pp. 44 164–44 164.
- [8] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, 2001.
- [9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr 1985.
- [10] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual