Advances in Intelligent Systems and Computing 1320

Anupam Biswas
Emile Wennekes
Tzung-Pei Hong
Alicja Wieczorkowska *Editors*

Advances in Speech and Music Technology

Proceedings of FRSM 2020

Springer

Advances in Intelligent Systems and Computing

Volume 1320

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Advisory Editors

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing, Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University, Gyor, Hungary

Vladik Kreinovich, Department of Computer Science, University of Texas at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro, Rio de Janeiro, Brazil

Ngoc Thanh Nguyen, Faculty of Computer Science and Management, Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within "Advances in Intelligent Systems and Computing" are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Indexed by DBLP, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST).
All books published in the series are submitted for consideration in Web of Science.

More information about this series at http://www.springer.com/series/11156

Anupam Biswas · Emile Wennekes ·
Tzung-Pei Hong · Alicja Wieczorkowska
Editors

Advances in Speech and Music Technology

Proceedings of FRSM 2020

Springer

Editors
Anupam Biswas
Department of Computer Science
and Engineering
National Institute of Technology Silchar
Cachar, Assam, India

Tzung-Pei Hong
Department of Computer Science
and Information Engineering
National University of Kaohsiung
Kaohsiung, Taiwan

Emile Wennekes
Department of Media and Culture Studies
Utrecht University
Utrecht, The Netherlands

Alicja Wieczorkowska
Multimedia Department
Polish-Japanese Academy of Information
Technology
Warsaw, Poland

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-33-6880-4 ISBN 978-981-33-6881-1 (eBook)
https://doi.org/10.1007/978-981-33-6881-1

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Conference Organization

Chief Patron

Sivaji Bandyopadhyay, Director, National Institute of Technology Silchar, India

Honorary Chair

Asoke Kumar Datta, Jadavpur University, India

Patron

Samir Kumar Borgohain, National Institute of Technology Silchar, India
Dipak Ghosh, Jadavpur University, India

General Chairs

Anupam Biswas, National Institute of Technology Silchar, India
Argha Deb, Jadavpur University, India

General Co-chairs

Pinki Roy, National Institute of Technology Silchar, India
Ranjan Sengupta, Jadavpur University, India

Organizing Chairs

Ripon Patgiri, National Institute of Technology Silchar, India
Shankha Sanyal, Jadavpur University, India

Program Chairs

Dalton Meitei Thounaojam, National Institute of Technology Silchar, India
Shymal Das Mandal, Indian Institute of Technology Kharagpur, India

Poster and Workshop Chairs

Thoudam Doren Singh, National Institute of Technology Silchar, India
Anish Kumar Saha, National Institute of Technology Silchar, India

Finance Chairs

Shyamosree Pal, National Institute of Technology Silchar, India
Malaya Dutta Borah, National Institute of Technology Silchar, India

Hospitality Chairs

Shyamapada Mukherjee, National Institute of Technology Silchar, India
Ujwala Baruah, National Institute of Technology Silchar, India
Suganya Devi K., National Institute of Technology Silchar, India
Laiphrakpam Dolendro Singh, National Institute of Technology Silchar, India

Publicity Chairs

Naresh Babu Muppalaneni, National Institute of Technology Silchar, India
Biswanath Dey, National Institute of Technology Silchar, India
Partha Pakray, National Institute of Technology Silchar, India

Local Committee Members

Biswajit Purakayashtha, National Institute of Technology Silchar, India
Arup Bhattachargee, National Institute of Technology Silchar, India
Saroj Kumar Biswas, National Institute of Technology Silchar, India
Badal Soni, National Institute of Technology Silchar, India
P. S. Neog, National Institute of Technology Silchar, India
Pantha Kanti Nath, National Institute of Technology Silchar, India
Umakanta Majhi, National Institute of Technology Silchar, India

International Advisory Committee

Suranjan Das, Vice Chancellor, Jadavpur University, India
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Solvi Ystad, Perception Representations Image Sound Music, CNRS, France
Xavier Serra, University of Barcelona, Spain
Kristoffer Jensen, University of Aalborg, Denmark
D. Duttamajumdar, President, SNLTR, Kolkata, India
John Ohala, University of California, Berkeley, USA
Aditi Lahiri, Oxford University, UK
M. Pushpavathi, Director, All India Institute of Speech and Hearing, Mysore, India
Hiroya Fujisaki, University of Tokyo, Japan
Anupam Basu, Director, National Institute of Technology Durgapur, India
Phalguni Gupta, Director, NITTTR, Kolkata
Keikichi Hirose, Emeritus Professor, University of Tokyo, Japan
J. Sundberg, KTH Royal Institute of Technology, Sweden
S. S. Agrawal, KIIT Group of Colleges, Gurgaon, India
Sanghamitra Mohanty, Former Vice Chancellor, North Orissa University, India
W. M. Hartman, Michigan State University, USA
Partha Ghose, S. N. Bose National Centre for Basic Sciences, Kolkata, India
S. G. Deshmukh, Director, Atal Bihari Vajpayee-Indian Institute of Information Technology and Management, Gwalior, India
Bhaskar Gupta, Jadavpur University, India
Javed Hashmi, Joint Secretary, Allenhouse Institute of Technology, Kanpur, India
Carlos Guedes, New York University Abu Dhabi, United Arab Emirates
Sanjay Kumar Singh, Indian Institute of Technology (BHU) Varanasi, India

Keynote Speakers

Keikichi Hirose, Emeritus Professor, University of Tokyo, Japan
Bjorn Schuller, Technical University of Munich, Germany
Michał Kosiedowski, Poznań Supercomputing and Networking Center, Poland
Sanghamitra Mohanty, Former Vice Chancellor, North Orissa University, India
T. V. Sairam, President, Nada Centre of Music Therapy, India
Aurobinda Rautray, Indian Institute of Technology Kharagpur, India

Session Chairs

Emile Wennekes, Chair Professor, Utrecht University, The Netherlands
Pradip Kumar Das, Indian Institute of Technology Guwahati, India
Sanjay Kumar Singh, Indian Institute of Technology (BHU) Varanasi, India
Debashis De, Maulana Abul Kalam Azad University of Technology, India
Sanghamitra Mohanty, Former Vice Chancellor, North Orissa University, India
K. L. Baishnab, National Institute of Technology Silchar, India
Binoy Krishna Roy, National Institute of Technology Silchar, India
Nidul Sinha, National Institute of Technology Silchar, India

International Program Committee

Mireille Besson, Research Director, CNRS, Aix-Mairseille University, France
Kildeep K. Paliwal, Griffith University, Australia
Alicja Wieczorkowska, Polish-Japanese Academy of Information Technology, Warsaw, Poland
Shahanum Mohd Shah, Universiti Teknologi MARA (UiTM), Malaysia
Phalguni Gupta, Director, NITTTR Kolkata, India
Sanghamitra Mohanty, Former Vice Chancellor, North Orissa University, India
George Tzanetakis, University of Victoria, Canada
Monojit Choudhury, Microsoft Research Lab, India
Adam Ockelford, Director, Applied Music Research Centre, University of Roehampton, UK
Debashis De, Maulana Abul Kalam Azad University of Technology, India
Pabitra Pal Chowdhury, Indian Statistical Institute Kolkata, India
Pradip Kumar Das, Indian Institute of Technology Guwahati, India
Asiya Khan, Plymouth University, UK
Bhaskar Biswas, Indian Institute Technology (BHU) Varanasi, India
Emile Wennekes, Chair Professor, Utrecht University, The Netherlands
Ajay Srinivas Murthy, Amazon, India

Kaustav Kanti Ganguly, New York University, Abu Dhabi
C. T. Bhunia, Former Director, NIT Arunachal Pradesh, India
Ferdous Ahmed Barbhuiya, Indian Institute of Information Technology Guwahati, India
Birjodh Tiwana, LinkedIn, San Francisco, USA
Nityananda Sharma, Tezpur University, India
Vijay Bhaskar Semwal, Maulana Azad National Institute of Technology Bhopal, India
Alex Noel Joseph Raj, Shantou University, China
Parismita Sarma, Gauhati University, India
Swanirbhar Majumder, Tripura Central University, India
Hima Bindu Maringanti, North Orissa University, India
Omar Farooq, Aligarh Muslim University, India
K. Samudravijay, Indian Institute of Technology Guwahati, India
Sauhaarda Chowdhuri, Resoniq Research, San Diego, USA
Subhendu Ghosh, Delhi University, India
Debasis Gangopadhyay, RKMVU, Belur, India
Sudhakar Tripathi, REC, Lucknow, India
Michał Kosiedowski, Poznań Supercomputing and Networking Centre, Poland
Josef van Genabith, University of Saarland, Germany
Alexander Gelbukh, Instituto Politecnico Nacional, Mexico
Mridula Verma, IDRBT Hyderabad, India
N. Parimala, Jawaharlal Nehru University, India
Debasis Das, Indian Institute of Technology Jodhpur, India
Rajni Jindal, Delhi Technical University, India
R. K. Agrawal, Jawaharlal Nehru University, India
Priyadarshi Patnaik, Indian Institute of Technology Kharagpur, India
Vibhav Prakash Singh, MNNIT Allahabad, India
Alain Trémeau, Université Jean Monnet, France
N. Sreedevi, AIISH, Mysore, India
Nandini Chatterjee Singh, NBRC Gurgaon, India
Denis Trcek, University of Ljubljana, Slovenia
Punita Singh, CSIR-NPL, New Delhi, India
Sanasam Ranbir Singh, Indian Institute of Technology Guwahati, India
Punam Kumar Saha, The University of Iowa, USA
Lairenlakpam Jaiprakash, North Eastern Hill University, India
Paolo Rosso, Universitat Politècnica de València, Spain
Shantala Hegde, National Institute of Mental Health and Neuro Sciences, Bengaluru, India
Archi Banerjee, Indian Institute of Technology Kharagpur, India

External Reviewers

Preeti Rao
Pradeep Kumar Roy
Dilip Kumar Choubey
Bhawana Rudra
Arka Banerjee
Vipin Chandra Pal
Durgesh Singh
Sayan Nag
Rajnish Pandey
R. Murugan
Rahul Chandra Kushwaha
Vishal Rohilla
Anand Bihari
Brijendra Gupta
Sachin Jain
Uddalok Sarkar
Avadh Pati
Banee Bandana Das
Rabi Shaw
Manish Mandloi
Vishwanath Bijalwan
S. D. Kamble
Madhusudhan Mishra
Amulya Sri Pulijala
Joyanta Basu
Akash Maity
Somi Kolita
Sapan Mankad
Swarnendu Moitra
V. Priyadharshini
S. V. Narasimhan
Subarna Chatterjee
Dipankar Das
Ganesh Chandra Deka
Samarjit Roy

Keynote Talks

Speech Synthesis and Prosody

Keikichi Hirose
University of Tokyo, Japan
hirose@gavo.t.u-tokyo.ac.jp

Abstract Prosodic features of speech play important roles in human communication, e.g., highlighting important words and expressing speaker's intentions or emotions, chunking and turn-taking. Since the current speech synthesis aims at increasing expressiveness and adding variances (i.e., various utterance styles) in generated speech, having control of prosodic features becomes very important. During the talk, after a brief summary on prosodic features with the well-known generation process model of fundamental frequency contours, history of speech synthesis is reviewed with an emphasis on prosody control. Then, I will zoom in treatments of prosody in the current speech synthesis technologies, e.g., HMM-based and DNN-based speech synthesis, and point out their problem, especially on "flexible" control of prosody. Generation process model is effective to tackle the problem. Some of our former works are introduced from this perspective—prosodic focusing (Fig. 1) and voice conversion. Finally, looking ahead, I plan to discuss how to realize various styles with less sized speech corpus.

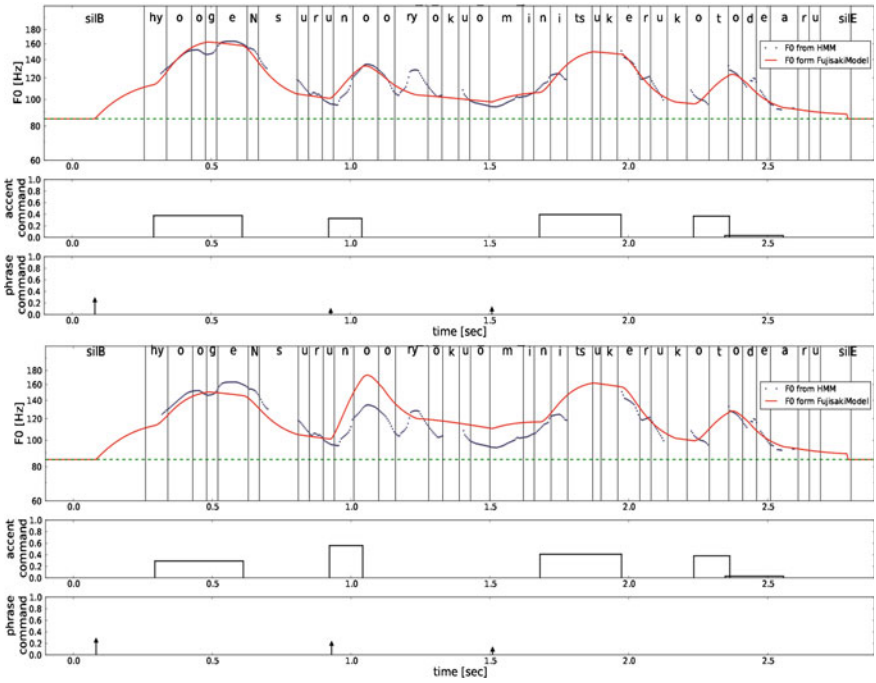


Fig. 1 Fundamental frequency contour generated by HMM-based speech synthesis (black line) and reshaped one using the generation process model (red line, upper panel). Word "nooryoku" (ability) is emphasized in the red line contour of the lower panel

Computational Paralinguistics: Challenge Accepted

Björn W. Schuller
Imperial College London, UK
schuller@informatik.uni-augsburg.de

Abstract Frontiers in research on computational paralinguistics—the automatic processing of speaker states and traits—are about to reach the point of broad everyday applicability of the technology. Over the last "two plus" decades, the field has massively widened up regarding targeted speaker information. Beyond the ID, gender and emotion, these days, biological and personality traits, health information and other rich information of speaker profiling become accessible to computing devices at increasing robustness. Here, we will touch upon the latest developments in and future trajectories of the field broken down into four major steps on the road to full applicability. Namely, these are (I) self-learning of representations, largely covered, e.g., by CNNs, or sequence-to-sequence AEs such as by LSTM-RNNs, potentially with attention mechanisms, (II) coupling of analysis and synthesis for data efficiency and mutual gains, increasingly covered by GANs or VAEs, (III) highly autonomous holistic big data learning in the wild, by reinforced, cooperative and self-supervised

learning, still largely lacking apart from the first step such as into Auto-ML for paralinguistics and (IV) assuring applicability factors such as real-world robustness including package loss concealment, de-noising, and de-reverberation, green efficient processing, explainable decision making, secure and private data handling, such as by federated learning and robustness against adversarial attacks, and usability and system integration aspects. In an outlook, we will discuss the advent of silent paralinguistics, and covering and general conversion of paralinguistic information. Usage of computational paralinguistics is around the corner—we need to be prepared on all ends—from technological to ethical aspects.

Extractive Summarization of Continuous Speech using Odia ASR

Sanghamitra Mohanty
Former Vice Chancellor, North Orissa University, India
sanghamitramohanty1@gmail.com

Abstract In this paper, an extractive speech summarization is handled by using HMM-based Odia Automatic Speech Recognition (ASR) engine. The continuous speech of meeting domain is used for the training of Odia ASR. The output of ASR engine is passed to the text summarizer module. The summarizer module arranges the whole ASR output into multiple sentences, and then, score of each sentence is calculated. Finally, the summarization module arranges the sentences by considering the highest score carrying sentences in non-decreasing order. A threshold value is used for limiting the contents of summarization result. Finally, a concatenative text-to-speech synthesis is used for producing the summarization result in voice-based synthesizer system.

Using Advanced ICT for Supporting People with Profound Intellectual and Multiple Disabilities to "Speak" to Others

Michał Kosiedowski
Head of Future Internet Services Department
Poznań Supercomputing and Networking Center, Poland
kat@man.poznan.pl

Abstract People with profound intellectual and multiple disabilities (PIMD) often communicate on the pre-symbolic level and are usually unable to use a verbal language. They interact with others using unconventional behavioral signals such as gestures, facial expressions and vocalizations. These signals are highly individual and are usually known only to a very limited number of familiar direct support

persons, e.g., parents and teachers. This, together with the level of the impairment, makes it extremely difficult for these people to engage in interaction with any other people and to influence their surroundings at will. However, the recent advances in such technologies as computer vision and artificial intelligence have created the ground for investigating whether ICT could be used to construct a sort of prosthesis of verbal communication for a person who is biologically unable to use verbal communication. This talk will take a look at what the potential of such technologies is to enable people with PIMD to gain a level of self-determination and independently act changing their surrounding as needed. The discussion of the topic will be based on the example of the research conducted within the Insension project, financed by European Union's Horizon 2020 program.

The Advantage of "Raga Chikitsa" in Modern Music Therapy

T. V. Sairam
President, Nada Centre of Music Therapy, India
tvsairam@gmail.com

Abstract We all know that culture and behavior are closely interconnected. This is for the first time that Dr. Sairam has come out with a definitive correlation between a musical culture and the behavior of its humans. Western system thriving on equal temperament scale and an elaborate notation system binds not only the performers but also its listeners to be meticulous and finicky. Adjusting A above C mechanically at 440 Hz was compared with "no-notation" manodharma rendering in the state of relaxed alertness as in Indian classical culture. Little or no instrument, always human voice-based, also proves further its "ordinariness" (not the "orderliness" of the West). West follows power and royalty of sound, while Indian music is simple, self-effacing, content, compassionate and inward-looking like yoga meditation. West follows the chordal progression at times with 100-piece orchestration which creates awe and extroversion, if not a "colonizing" aggression.

An Application for Early Warning of Conflict and Aggression in Dyadic Speech Communication

Aurobinda Rautray
Indian Institute of Technology Kharagpur, India
aurobinda.routray@gmail.com

Abstract In social interactions, aggression is a behavior toward another individual with the motive of causing physical or psychological damage. In humans, anger or disgust due to obstructions in achieving certain goals causes aggression. This talk shall discuss an automatic method for detection of aggression using features extracted from pressure distribution in the vocal tract during voiced speech. The variations in air pressure distribution across different sections of the vocal tract have been computed from the speech signal by inverse estimation. A hidden Markov model has been trained to classify these air pressure variations as aggression or calm. The system has been tested on a set of audio clips extracted from interviews of political personalities on television shows. A total of 120 audio clips with an overall duration of around 30 min were collected from three different speakers based on human perception of aggression and calm. The clips were rated by human raters to discard perceptually ambiguous clips. The speaker-dependent system was trained using 40% of the data of each speaker and was tested with the remaining 60% of the data of the same speaker. The system was able to detect aggression with about 93% accuracy.

Preface

FRSM 2020 was the Silver Jubilee Edition of International Symposium on Frontiers of Research in Speech and Music (FRSM) organized jointly by the Department of Computer Science and Engineering, National Institute of Technology Silchar and Sir C. V. Raman Centre for Physics and Music, Jadavpur University. This unique annual international event intends to bring researchers in the field of speech and music together. At the same time, it also intends to pitch the objective and cognitive research potentials of these two pre-historic fields of human communication to researchers in different disciplines including those of science and technology. The 25th edition of FRSM 2020 was held virtually during October 08–09, 2020, and hosted by the National Institute of Technology Silchar, India.

The FRSM is a well-established symposium organized in different parts of India every year since 1990. From 1990 to 2001, it was known as International Workshop on Speech, Music and Allied Signal Processing (IWSMSP). In 2003 (at IIT Kanpur), it changed its name to the present form, i.e., Frontiers of Research on Speech and Music (FRSM). There has been a gap of certain years because of unavoidable circumstances. Until now, these symposiums were held at Kolkata, Santiniketan, New Delhi, Varanasi, Kanpur, Lucknow, Mysore, Gwalior, Bhubaneswar, Annamalai, Kharagpur, Baripada, Thiruvananthapuram and Rourkela.

The continuation of these events would not have been possible without the active cooperation of ITC Sangeet Research Academy, Kolkata, SCVRCPM-Jadavpur University, Kolkata, CDAC-Kolkata, Indian Statistical Institute, Central Electronics Engineering Research Institute, BHU, Visva-Bharati University, International School of Dravidian Linguistics, IIT Kanpur, Annamalai University, Utkal University, Uttar Pradesh Technical University, IITM Gwalior, All India Institute of Speech and Hearing, Mysore, IIT Kharagpur, NOU Baripada, NIT Rourkela, AIT Kanpur, Acoustical Society of India and active participation of various research groups, particularly Tata Institute of Fundamental Research, Indian Institute of Science, various IITs, and Indian and foreign universities and institutions.

Although the FRSM is an international event, earlier editions of FRSM were mostly focused only on Indian participants working on Indian classical music and speech. On the 25th edition of FRSM, i.e., FRSM 2020, participants from other

parts of the world also presented their research findings. We received papers from many countries all over the world. Each paper was peer-reviewed by at least three members of the international program committee and external reviewers, and most of the papers were reviewed by 5–7 reviewers. Only 36 papers of the highest quality were selected for oral presentation and publication in the proceedings of FRSM 2020.

The papers included in the proceedings are organized into five parts, covering both technological advancement and interdisciplinary nature of speech and music processing. The first part contains chapters covering the foundations of both vocal and instrumental music processing. The second part includes chapters related to computational techniques involved in the speech and music domain. A lot of research is being performed within the music information retrieval domain which is potentially interesting for most users of computers and the Internet. Therefore, the third part is dedicated to music information retrieval. The fourth part contains chapters on the brain signal analysis and human cognition or perception of speech and music. The final part consists of chapters on spoken language processing and applications of speech processing.

We would like to extend our heartfelt thanks to Prof. Sivaji Bandyopadhyay, the Director of National Institute of Technology Silchar, for his support and honorary patronage over the symposium. We also thank the co-patrons, Dr. Samir Kumar Borgohain (HOD, Department of Computer Science and Engineering, National Institute of Technology Silchar) and Prof. Dipak Ghosh (Jadavpur University) for their support.

We would like to express our sincere thanks to our honorary chair, Prof. A. K. Dutta (Jadavpur University), and general chairs, Dr. Anupam Biswas (National Institute of Technology Silchar) and Prof. Argha Deb (Jadavpur University), for their support and guidance.

Our special thanks go to the general co-chairs, organizing chairs, program chairs, session chairs, poster and workshop chairs, finance chairs, hospitality chairs, publicity chairs and local committee members for their commendable work for the symposium. We sincerely thank all the international program committee members and external reviewers for their valuable efforts in the review process, which helped us to guarantee the high-quality papers for the symposium. A special word of thanks goes to the intern from Utrecht University, Leroy Roncken, who assisted in the editing process.

We would like to express our thanks to the keynote speakers, Prof. Keikichi Hirose (Emeritus Professor, University of Tokyo, Japan), Prof. Bjorn Schuller (Imperial College London, UK), Prof. Sanghamitra Mohanty (Utkal University and former VC of North Orissa University, India), Dr. Michał Kosiedowski (Poznań Supercomputing and Networking Center, Poland), Dr. T. V. Sairam (President, Nada Centre of Music Therapy, India) and Prof. Aurobinda Rautray (Indian Institute of Technology Kharagpur, India) for their informative talks.

We cordially thank all the authors for their valuable contributions. The symposium would not have been possible without their research inputs.

Lastly, we thank all the stakeholders who have contributed directly or indirectly for making the event a success.

Silchar, India — Anupam Biswas
Utrecht, The Netherlands — Emile Wennekes
Kaohsiung, Taiwan — Tzung-Pei Hong
Warsaw, Poland — Alicja Wieczorkowska
November 2020

Contents

Vocal and Instrumental Music Processing

Music Signal Processing: A Literature Survey 3
Amulya Sri Pulijala and Suryakanth V. Gangashetty

Effects of Vocal Loading on Singing Power Ratio and Singer's Formant with Indian Heavy Metal Vocalists 17
Mahesh Soumya and S. V. Narasimhan

Audio Quality Control Method Based on ASS (Audio Secret Sharing) 25
Sachin Kumar Singh and Mainejar Yadav

Noise Removal from Audio Using CNN and Denoiser 37
Manmohan Dogra, Saumya Borwankar, and Jayashree Domala

Sine-Wave Speech as Preprocessing for Downstream Tasks 49
Sri Harsha Dumpala, Jason d'Eon, and Sageev Oore

Style of Vocal Singers in Indian Classical Music: Timbre Approach 55
Anirban Patranabis, Kaushik Banerjee, Ranjan Sengupta, and Dipak Ghosh

Style Identification of Vocal Singers in Indian Classical Music Using Meend and Andolan 67
Kaushik Banerjee, Anirban Patranabis, Ranjan Sengupta, and Dipak Ghosh

Computational Music and Speech

Speech Based Emotion Detection Using R 85
Nida Farheen and Subarna Chatterjee

Vocalist Identification in Audio Songs Using Convolutional Neural Network 99
Manoj Kaushik, Shikha Rani, and Vrinda Yadav

Swaragram: Shruti-Based Chromagram for Indian Classical Music 109
Yeshwant Singh, Ranjeet Kumar, and Anupam Biswas

Machine Learning Approach for Audio Surveillance Using R 119
Nida Farheen and Subarna Chatterjee

An Artificial Intelligence-Based Approach Towards Segregation of Folk Songs 133
Himadri Mukherjee, Ankita Dhar, Sk. Md. Obaidullah, Santanu Phadikar, Kaushik Roy, and Sukumar Nandi

Shruti Detection Using Machine Learning and Sargam Identification for Instrumental Audio 145
Muktharunisa Chapparband, Megha Gururaj Kulkarni, D. K. Sameeksha, A. Vijaya Krishna, and Ashwini Bhat

Addressing the Recitative Problem in Real-Time Opera Tracking 157
Charles Brazier and Gerhard Widmer

Music Information Retrieval

IFSC: A Database for Indian Folk Songs Classification 171
Anshul Patel, Anuj Shah, Krutarth Gor, and Sapan H. Mankad

North Indian Classical Music Tabla Tala (Rhythm) Prediction System Using Machine Learning 187
Shambhavi Shete and Saurabh Deshmukh

Perception of Similarity and Dissimilarity in Hindustani Classical Music 199
Junmoni Borgohain, Gouri Karambelkar, Raju Mullick, and Priyadarshi Patnaik

Analytical Comparison of Classification Models for Raga Identification in Carnatic Classical Audio 211
Sathwik Acharya, Vihar Devalla, Om Amitesh, and Ashwini

Multimodal Sentiment Analysis of Rabindra Sangeet Through Machine Learning Techniques 223
Sandipan Ganguly, Dipankar Das, Akash Modak, and Samanway Chakraborty

Music Genre Classification with Convolutional Neural Networks and Comparison with F, Q, and Mel Spectrogram-Based Images 235
Aditya Dhall, Y. V. Srinivasa Murthy, and Shashidhar G. Koolagudi

A Simple Statistical Comparison of Bageshree and Bhimpalasi 249
Pranjala Shukla and Soubhik Chakraborty

Sound Analysis and Cognition

Effects of Musical Training on Auditory Spatial Processing Abilities: A Psychoacoustical and Perceptual Study 261
Bhoomika and Kavassery Venkateswaran Nisha

Effect of Transformation of Vowel Duration on Speech Intelligibility in Children with Hearing Loss 275
S. V. Narasimhan and N. P. Nataraja

The Influence of Music on Image Making: An Exploration of Intermediality Between Music Interpretation and Figurative Representation 285
Pinaki Gayen, Junmoni Borgohain, and Priyadarshi Patnaik

Comparison of Neural Network Architectures for Speech Emotion Recognition 295
Shubham Patil and Saaket Agashe

A Comprehensive Review on Analysis Methods, Software, and Application of fMRI for Classification of Alzheimer's Disease 309
Chintan R. Varnagar and Hitesh B. Shah

A Pilot Study of Indian Ragas-Based Music Therapy for Enhancement of Chronic Patient Health Condition 325
Milind U. Nemade

Scottish Education: A Model for Strengthening Mental Health—A Case Study in Inclusive Scottish Primary Schools 341
Hanah Kim

Spoken Language Processing

Language Model Fine-Tuning with Second-Order Optimizer 355
Kartik Vyas and Dinesh Naik

Human-to-Robotic Voice Conversion by Employing Wave Patterns Based on Gender and Emotions 367
Dipankar Das and Rubel Das

Classifying Hate Speeches Shared in Twitter 381
Dipankar Das, Souvik Mondal, and Antariksha Ray

Impact of Visual Representation of Audio Signals for Indian Language Identification 395
Himanish Shekhar Das and Pinki Roy

Bi-Lingual TDNN-LSTM Acoustic Modeling for Limited Resource Hindi and Marathi Language ASR 409
Ankit Kumar and Rajesh Kumar Aggarwal

Vowel-Based Acoustic and Prosodic Study of Three Manipuri Dialects 425
Thangjam Clarinda Devi, Leihaorambam Sarbajit Singh, and Kabita Thaoroijam

A Continuous Speech Recognition System for Bangla Language 435
Rajat Paul and K. Samudravijaya

Indian Language Speech Label (ILSL): A De Facto National Standard 449
K. Samudravijaya

Author Index 461

About the Editors

Dr. Anupam Biswas received his Ph.D. degree in computer science and engineering from Indian Institute of Technology (BHU), Varanasi, India, in 2017. He has received his M.Tech. and B.E. degrees in computer science and engineering from Nehru National Institute of Technology Allahabad, Prayagraj, India, in 2013, and Jorhat Engineering College, Jorhat, Assam, in 2011, respectively. He is currently working as Assistant Professor in the Department of Computer Science and Engineering, National Institute of Technology Silchar, Assam. He has published several research papers in reputed international journals, conference, and book chapters. His research interests include computational music, machine learning, fuzzy systems, information retrieval, and evolutionary computation. He has served as Program Chair of International Conference on Big Data, Machine Learning and Applications (BigDML 2019). He has served as General Chair of 25th International Symposium Frontiers of Research in Speech and Music (FRSM 2020).

Prof. Dr. Emile Wennekes is Chair Professor of Musicology at Utrecht University, The Netherlands. He was appointed full professor there in 2000. From 2006 to 2011, he was the first Head of the School of the Media and Culture Studies department. In 2017, his chair was modified from Musicology: Post-1800 Music History into Musicology: Music and Media, now also officially embracing his main field of research. Wennekes has written on a broad range of subjects, including a co-authored book on contemporary Dutch music available in six languages. His work has been published by, among others, Oxford University Press, Routledge, Michigan University Press, and Brepols. Most recently, he edited the volume Cinema Changes: Incorporation of Jazz in the Film Soundtrack (Brepols 2019) together with Dr. Emilio Audissino. Wennekes founded and chairs the Study Group Music and Media (MaM) under the auspices of the International Musicological Society. He coordinates its annual conferences.

Tzung-Pei Hong received his Ph.D. degree in computer science and information engineering from National Chiao Tung University in 1992. He was Associate Professor in Chung Hua Polytechnic Institute from 1992 to 1994 and in I-Shou University from 1994 to 1999. He served as the first director of the Library and Computer Center, Dean of Academic Affairs, Administrative Vice President, and Academic Vice President in National University of Kaohsiung. He is currently Chair Professor at the Department of Computer Science and Information Engineering. He has published more than 600 research papers in international/national journals and conferences and has planned more than fifty information systems. His current research interests include data mining, artificial intelligence, soft computing, and granular computing.

Alicja Wieczorkowska Ph.D., D.Sc., is a computer scientist, specializing in audio signal analysis. She holds a Ph.D. from the Gdansk University of Technology, and additionally she is the alumna of the State School of Music (Second Level). Her Ph.D. thesis examined the automatic recognition of musical instrument sounds, depending on parameterization and classifiers applied. Dr. A. Wieczorkowska is presently Associate Professor and Head of Multimedia Laboratory at the Polish-Japanese Academy of Information Technology, Warsaw, Poland. Additionally, she is also an associate member of the Graduate Faculty at the University of North Carolina at Charlotte. Her scientific interests include audio information retrieval, data mining, as well as computer graphics, multimedia, and automated identification of emotions from various signals.

Vocal and Instrumental Music Processing

Timbre, Instrumentation and Voice, Rhythm, Beat and Tempo, Noise Reduction, Segmentation, Transcription, Musical Style and Genre

Music Signal Processing: A Literature Survey

Amulya Sri Pulijala and Suryakanth V. Gangashetty

Abstract Music has become an integral part of our lives today. With the digital revolution that has struck the world and with the growth of computational power, browsing and storage have become accessible and effective. Thus, audio signal processing became an emerging area, paving the way for many new areas of research. In this paper, an attempt is made to give an overview of existing areas of research in music signal processing. Existing methodologies in these respective areas are explained in detail. A brief overview of future perspectives is also discussed.

Keywords Music signal processing · Music information retrieval · Carnatic music · Hindustani music · Raga recognition · Tala classification · Melody/intonation analysis · Source separation

1 Introduction

Music signal processing is a sub-branch in signal processing, which is a promising area these days. Music analysis based on signal processing techniques paved a new way for generation and analysis of music. Music analysis and recognition of various swaras and ragas is germane to human understanding. Music signal processing includes various areas of research, such as synthesis of music, transcription, classification, music information retrieval, raga classification, tala classification, instrument/voice identification, audio matching, source separation, tonic identification, intonation/melodic/rhythmic analysis, and music emotion recognition. This paper gives an overview of existing areas of music signal processing primarily focusing on Indian Carnatic music.

A. S. Pulijala (✉) · S. V. Gangashetty
International Institute of Information Technology, Hyderabad, India

S. V. Gangashetty
e-mail: svg@iiit.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_1

1.1 Indian Classical Music

The oldest available books with respect to Indian Classical music are Bharata's Natya Shastra and Saranga Deva's Sangita Ratnakara [1]. These comprise 22 musical notes (practically only 16 are useful). The primary aspects of Carnatic music are swara, raga, tala, and tonic aspects of raga and shruti. Indian Classical music can be classified as Hindustani or Carnatic based on its origin. Hindustani music is mostly performed in the northern part of the country, while Carnatic belongs to the southern region.

Indian Classical music has seven swaras known as Saptha swaras. They are Shadjamam (Sa), which is the base frequency or tonic frequency, Rishabam (Ri), Gandharam (Ga), Madhyamam (Ma), Panchamam (Pa), which is the perfect fifth, Dhaivatham (Dha), and Nishadam (Ni). Of these, Sa and Pa do not admit variations and are called achala (fixed) swaras [1]. The other five (Ri, Ga, Ma, Dha, and Ni) have variations in them and are indicated by numbers 1, 2, and 3, thus corresponding to the total of 16 notes, of which. Ri3, Ga1, Da3, and Ni1 are hardly used.

There is a mythological origin of swaras that tells us that Lord Shiva addressed a congregation—the differences that existed in His tone became the seven swaras of music. Shiva first addressed the audience at the center and then addressed the audiences to the immediate left and right. The center tone became the basic note, or the Sadjamam (Sa), while the swaras on the immediate left became the ni, and the right became the ri. The Lord then addressed the far left and right. These became the notes dha, ga, and ma. Thus, seven swaras: Sa, Re, Ga, Ma, Pa, Dha, and Ni were formed [2]. The tonal quality of each note in the seven swaras is often associated with the call of specific animals, shown in Table 1.

Raga

One of the most important elements of Carnatic music is raga. It can be described as the spine of Indian Classical music. The following is the definition of "raga": "Ranjayati iti Raaga":

Translation: That which pleases the ear is described as raga. Matanga, in his epic treatise Brihaddeshi, deems raga as that which colors the mind of good, through a specic swara1 and varna (meaning color), or through a type of dhvani (meaning

Table 1 Swaras and their associated animal sounds

Name of the swara	Notation	Swara
Sadjamam	Sa	Cry of the peacock
Rishabham	Ri	Lowing of the bull
Gandhara	Ga	Bleating of the goat
Madhyamam	Ma	Call of the heron
Panchamam	Pa	Call of the cuckoo
Dhaivata	Dha	Neighing of the horse
Nishada	Ni	Trumpeting of the elephant

sound). Ragas are broadly classified as Melakartha and Janya ragas or sampoorna and asampoorna ragas. A sampoorna raga includes all the seven swaras in both the aarohana and the avarohana sequence. Any other raga which does not consist of all seven swaras is asampoorna. All Melakartha ragas are sampoorna ragas, while Janya ragas may be either sampoorna or asampoorna ragas. Another characteristic of the Melakartha ragas is that both the aarohana and avarohana sequences are required to contain exactly one note from each swara. And in the aarohana sequence, the pitch of the next swara must always be higher than the previous swara. However, Janya ragas can contain seven notes, more, or less. Aarohana and avarohana of Janya ragas can have repetitions [1].

Tala

Tala is nothing but the strong tradition of rhythm. In Carnatic music, we have many talas, namely Adi Talam, Rupaka Talam, etc. The combination of the first letters of Tandava and Lasya gives us the word tala: Tandava is the dance of Lord Shiva, and Lasya is the dance of his consort Shakti. Thus, tala is a combination of energy and grace [3].

Tala has no reference in the earliest system of music, popularly referred to as "Samagana". The Natyasastra of Bharata and the Sangitaratnakara of Sarngadeva are the oldest existing sources, as mentioned above, and are the main sources of information on tala.

A tala is characterized by ten features; hence, it is called the Tala Dasapranas (Dasa—ten, Prana—essence) [4]. The following are brief descriptions of these Dasapranas:

- Anga: part or limb
- Jati: type or kind. It describes variations in Anga (Laghu)
- Kriya: action
- Kaala: duration or measurement of time
- Graha: describes where the song commences, may not be at the beginning of the tala
- Marga: path. Describes duration of kriya/action. In other words, how tala is performed in various different songs
- Kala: denotes number of matras in which kriya is subdivided
- Laya: time gap between two consecutive kriyas. It sets the tempo
- Yati: rhythmic pattern in composition with reference to anga
- Prasthara: detailed elaboration of rhythmic pattern.

Tala Schemes in Carnatic Music

Carnatic music has various schemes of talas. The ancient 108 Anga Talas (which include the five Margi and 103 Desi Talas), the 72 Melakartha Tala system (which was designed to fit the 72 Melakartha Raaga classification), and the Suladi Sapta Tala system are some of the classification schemes.

Saptha Tala System

The Saptha Tala system has become very famous during the Purandaradasa period. As explained earlier, a tala is a cyclic repetition of a given rhythmic pattern. An Avartanam is a complete cycle of talas which consists of Aksharas, which are further divided into Mathras. The seven talas are as follows: Dhruva, Matya, Rupaka, Jhampa, Triputa, Ata, and Eka talam. These seven talas are subdivided based on the change in tala due to a change in five jaathis (jaathis of tala means the amount of beats that a laghu can take). The five jaathis are as follows: Tisra, Chatusra, Khanda, Misra, and Sankeerna. Thus, we get a total of 35 talas after dividing on the basis of jathis. These 35 talas allow further subdivision based on five gathis (or nadais) (gathis means speed). The five gathis are the same as the jaathis above. Finally, after jathi and gathi subdivision of seven talas, we get a total of 175 talas in Carnatic music. Three elements, namely the jaathi name, the tala name, and the gati name, are required to describe a tala [5].

Tonic

The tonic, or sruthi, is a fundamental component in Indian Classical music. It is the pitch chosen by the performer which is constant throughout the performance and acts as the reference frequency [6]. Swara Sa in octave is considered as the tonic/sruthi. Generally, instruments like tambura, etc., are played to establish the tonic. Accompanying instruments also tunes to the specific tonic of the performer. Scientifically, it is nothing but the linear shift on a time–frequency graph. Tonic identification can be done either by tuning with a drone (tambura) or melodic characteristics. With this introduction of basic elements of Carnatic music, let us look at the elements of a Carnatic concert and the instruments used in the concert, which are explained in detail in the consecutive sections.

1.2 Carnatic Concert

In any Carnatic music concert, the emphasis is always on the vocal music and the singer. From the perspective of the music listener, the Carnatic music concert is more like storytelling or the conversation between artist and audience. There is no definite structure; however, people conventionally start with a varnam followed by a kriti. The varnam is generally regarded as a warm-up artwork, which sets the pace and mood of the concert. This is usually followed by a kriti with or without raga alapanam. Then, the performer renders the main musical work of the concert in an elaborate and detailed fashion. Next comes a tani avartanam, where percussionists rhythmically embellish. Generally, most artists sing a Ragam Tanam Pallavi as the secondary major piece after the main piece. This is followed with a variety of songs, like Bhajans, Slokas, Ashtapadis, etc. The concert concludes with mangalam, which is a prayer for peace and prosperity.

1.3 Classification of Indian Musical Instruments

Erich Moritz von Hornbostel and Curt Sachs devised a system for classification of musical instruments, which is used by orgnalogists worldwide (people who study musical instruments) [7, 8]. According to the Hornbostel-Sachs system, Indian musical instruments can be classified as:

- Chordophones (string instruments)
- Aerophones (wind instruments)
- Membranophones (drum percussion instruments)
- Idiophones (non-drum percussion instruments).

Chordophones primarily produce their sounds by means of the vibration of a string or strings that are stretched between points. Examples of chordophones include the violin, sarangi, veena, and the sitar. Aerophones are the kind of instruments in which the air itself acts as the primary vibrating medium for the production of sound. Examples include the flute, harmonium, nadaswaram, and the ottu. The term "percussion instrument" refers to the method of playing the instrument, namely striking it either by fingers, hand, or by using sticks. This class of instruments is classified into two kinds, namely membranophones and idiophones. Idiophones are the kind of music instruments whose substance vibrates to produce sound. Membranophones produce sound by the vibration of a stretched membrane. Examples are the mrudangam, tabla, ghatam, khanjira, morsing, chanda, etc.

These categories are further classified, and each of them is explained in Fig. 1.

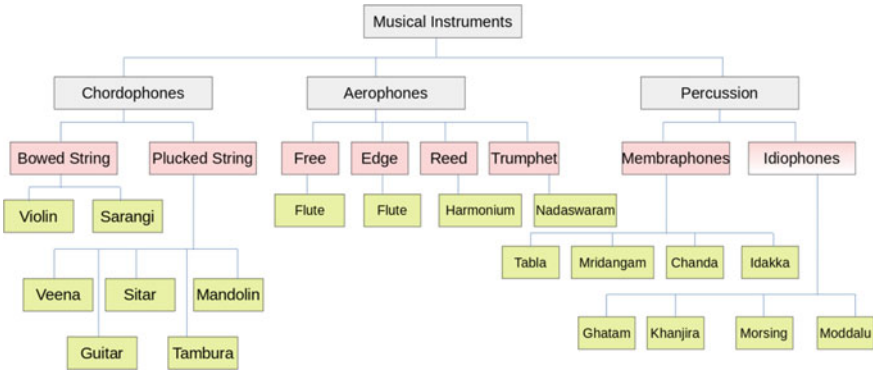


Fig. 1 Categories of musical instruments

2 Areas of Research

This section gives an overview of various analysis methods involved in music signal processing. The following are the broad areas that can be classified from the literature.

- Source Separation
- Tonic Identification
- Intonation/Rhythmic Analysis
- Ontologies
- Automatic Music Stretching Resistance
- Music Emotion Recognition
- Music Note Representation
- Raga Identification
- Tala Identification
- Music Therapy.

Source separation, or audio source separation, is the act of separating and isolating sound sources. There are many applications for source separation like voice separation, instrumental separation, karaoke extraction, etc. Various methods of source separation are described in the literature. Each of the methods is explained briefly in the consecutive paragraphs.

2.1 Source Separation

- Source Separation Based on Harmonic Structure: This method consists of three stages: (a) estimating the harmonic structure of each source in every frame based on iteration of the mixed spectral peaks (b) clustering the estimated harmonics into the signals they belong to with pitch and formant information, and (c) synthesizing the music source in the time domain. This method has the advantage of solving the octave overlapping problem [9].
- The Bayesian Harmonic Model: This estimates multiple fundamental frequencies in the frequency domain and combines the Bayesian harmonic model to estimate other parameters of the music signals. In the next stage, the algorithm separates all estimated parameters [10].
- Coherent modulation spectral filtering [11].
- The inter-channel level vector sum (ILVS) concept clusters common signals (such as background music) from each channel. The advantage of this method is that it separates audio signals from multiple channels with lower complexity [12].
- Autoregression models [13].
- Three strategies are used, namely Gaussian scale mixture models (GSMM), autoregressive (AR) models, and amplitude factor (AF) [14].
- Non-negative matrix factorization [15–18].
- Binary time frequency annotation [19].

- Mask optimization in a recurrent neural network (RNN): This method is used to assess the impact of reducing the binary masks in deep neural networks by averaging time and frequency bins, so that computational cost can be reduced. The ideal soft mask is then compared against masks produced by a range of averaging levels [20].
- Wiener filter modification: A finite impulse response (FIR) filter, which is suitable for distilling sound of a part played by a particular instrument from a music ensemble, is discussed in this report. The filter separates a monaural music signal into individual tones of every pitch with a particular timbre, and the combination of which provides melodies played by an instrument. Wiener filter is combined with a window and an evaluation method of current statistics of the mixture in the frequency domain [21–23].
- Spectrum inversion [24].
- Fourier transform [25, 26].
- Principal component analysis [27, 28].
- Repeating musical structure [29–31].

2.2 Emotion Recognition

Emotion recognition is another major area in music signal processing that has become popular in recent times. The paper "EEG-Based Emotion Recognition in Music Listening" describes a novel method for estimating the emotional responses elicited by a piece of music from its acoustic signals [32]. The focus of emotion recognition in recent days has been based on a multi-stage regression. After training, the aggregation happens in a predefined way and cannot be adapted to acoustic signals with different musical properties. It proposes a method that adapts the aggregation by taking into account new acoustic signal inputs. Another important way of carrying out emotion recognition is a system for detecting emotion in music that is based on a deep Gaussian process (GP) [33, 34].

The system can be classified as two stages, namely feature extraction and classification. In the first stage, rhythm, dynamics, timbre, pitch, and tonality are considered as features. The given musical piece is segregated into frames, and these five features are extracted for each frame. Next, various statistical parameters, such as mean and standard deviation, are calculated to generate a 38 feature vector. In the second stage, a deep GP is used in order to classify emotions into nine different classes.

2.3 Raga Classification

Classification and identification of ragas is a problem that has been addressed by several works of research in the recent past. Parag Chordia and Alex Rae have used the pitch class and the bigrams of pitches to classify ragas [35]. The dataset used

in the proposed system consists of 72 min of monophonic instrumental (sarod) data in 17 ragas played by a single artist. The harmonic-percussive separation (HPS) algorithm is used to extract the pitch. Note onsets are detected by observing the sudden changes in phase and amplitude in the signal spectrum. Then, the pitch class and the bigrams are calculated. It is shown that bigrams are useful in discriminating the ragas with the same scale. The proposed model uses several classifiers combined with dimensionality reduction techniques. Using just the pitch class, the system achieves an accuracy of 75%. Using only the bigrams of pitches, the accuracy is 82%. Rajeswari Sridhar and T.V. Geetha used an approach to identify ragas based on individual swaras that were identified in the music recording [36]. The frequency components extracted from the recording were matched with a database containing the swara sequences for each raga [37]. Another approach of raga classification uses locality sensitive hashing (LSH), which finds and matches similar data points hashed into various buckets [38]. Another approach uses longest common segment set (LCSS) [39].

Trupti Katte proposed a method which used the "pakad" to identify the raga of a musical piece [40]. The text "Phrase-based rāga recognition using vector space modeling" uses phrase-based raga identification [41].

2.4 Tala Classification

Literature [42] used both statistical as well as deep learning methods, and a two-dimensional convolutional neural network (CNN) gave the accuracy of 92%. The Inception V3 architecture of a CNN resulted in an accuracy of 50%. This was followed by kNN with k set to 3, which gave us an accuracy of 42%. Although statistical methods proved to give better accuracy, they lack robustness and a generalizing ability.

2.5 Intonation and Rhythmic Analysis

Intonation is an important characteristic of raga. Koduri et al. [43] proposed two approaches based on the parameterization of pitch-value distributions: Performance pitch histograms and context-based swara distributions were obtained by categorizing pitch contours based on the melodic context and evaluated both these approaches on a larger dataset of Carnatic music. Koduri et al. [44] proposed intonation from pitch track. The procedure is as follows: First, extract the pitch contours from n selected voice segments. Then, obtain a histogram normalized by the tonic frequency, from which, each prominent peak is automatically labeled and parameterized. Subramanian [45] analyzed intonation with respect to thodi raga.

Work by Guedes et al. [46] presented a data-driven approach for generating rhythmic analysis in Carnatic music. Guedes et al. [46] used arithmetic partitions

where each partition consists of a combination of stroke sequences. Ganguli and Guedes [47] added knowledge constraints to this data-driven approach. Heshi et al. [48] used rhythm and timbre patterns. Trochidis et al. [49] generated rhythm using n-gram Markov chains. Sebastian and Murthy [50] used separation-driven onset detection methods for detecting stroke locations of percussion instruments.

2.6 Tonic Identification

The tonic can be described as the base frequency of the singer. All the instruments used in the concert are adjusted to the tonic frequency of the singer. Normalizing the music recording with the tonic frequency is the first step in Carnatic music analysis. Gulati et al. [6] described the methodology of tonic pitch identification in two stages: In the first stage, which is applicable to both vocal and instrumental music, a multi-pitch analysis of the audio signal is performed to identify the tonic pitch class. In the second stage, the octave is estimated in which the tonic of the singer is thus needed only for the vocal performances. Bellur and Murthy [51] used group delay to identify the tonic frequency.

Samsekai Manjabhat et al. [52] proposed a two-stage process, where in the first phase, the tonic and raga have been determined independently using the features extracted from a pitch histogram. In the second phase, the raga and tonic are updated iteratively using the derived note information. Salamon et al. [53] used multi-pitch analysis for the same. Bellur et al. [54] use strategies like template matching, concert method, and segmented histogram to evaluate the tonic frequency.

2.7 Musical Note Representation

Mammen et al. [55] formed a unicode-based music notation representation language as an alternative to many ASCII representations, such as those given in [56] and [57]. Sridhar and Geetha [58] perform segmentation of Carnatic music and then find the pitch associated with respect to each segment. Prashanth and Venugopalan [59] attempted identification of swaras using statistical T-Test, while the authors of [60] proposed swara identification using stochastic models.

2.8 Music Therapy

Music therapy is a way of improving the quality of life using music and musical instruments. Music therapy has gained significant attention these days from both academics and industries. The book written by Bunt describes the process of music therapy and growth of it, especially in uplifting people's psychodynamic aspects [61].

Bunt and Stige [62] elaborate on how music therapy helps to cope with depression. Research supports the effectiveness of music therapy with autistic children [63].

3 Conclusion and Future Work

Music signal processing is a swiftly evolving area with a variety of challenges. This paper attempted to present a detailed overview of existing literature and methodologies in the area of music signal processing. However, the techniques mentioned above are not the end of music signal processing; there are many challenges and exciting problems left such as audio thumbnailing, audio matching, version identification, tala classification of all existing 175 talas, composer identification, album detection, automatic generation of ragas, and identification of song names. Human perception and recognition of music is far beyond machine understanding and extraction. We hope these developments continue progressively and lessen this gap between human and machine perception and make music more colorful.

References

1. Sharma M (2007) Musical Heritage of India. APH Publishing,
2. Mahesh A (2017) Origin of Swara. https://anuradhamahesh.wordpress.com/carnatic-music-lessons-2/15-mythological-origin-of-the-swara/ Accessed 3 June 2020
3. Shankar SV (2002) Tala Anubhava of the music trinity. J Music Acad.
4. Upbeat labs (2018) A primer for carnatic talas. Accessed 3 June 2020
5. Balaji MJ (2020) Introduction to Talas and Sapta Tala System, 2008. Accessed 3 June 2020
6. Gulati S, Salamon J, Serra X (2012) A two-stage approach for tonic identification in Indian art music. In: Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12–13; Istanbul, Turkey. Barcelona. Universitat Pompeu Fabra
7. Kartomi MJ (1990) On concepts and classifications of musical instruments. University of Chicago Press, Chicago
8. Von Hornbostel EM, Sachs C (1961) Classification of musical instruments: translated from the original german by anthony baines and klaus p. wachsmann. Galpin Soc J
9. Wang D, Huang Q (2009) Single channel music source separation based on harmonic structure estimation, pp 848–851. https://doi.org/10.1109/ISCAS.2009.5117889
10. Wang H, Wang Y, Wang W, Zhu B, Ma S (2011) Single channel polyphonic music signal separation based on Bayesian harmonic model. In: Proceedings of 4th international congress on image and signal processing, Shanghai, pp 2784–2787. https://doi.org/10.1109/CISP.2011.6100778
11. Atlas L, Janssen C (2005) Coherent modulation spectral filtering for single-channel music source separation. In: Proceedings (ICASSP'05). IEEE international conference on acoustics, speech, and signal processing, 2005. vol 4. IEEE (2005)
12. Park T, Kang KO (2014) Background music separation for multichannel audio based on Interchannel level vector sum. In: Proceedings of 18th IEEE international symposium on consumer electronics (ISCE 2014), JeJu Island, pp 1–2. https://doi.org/10.1109/ISCE.2014.6884340
13. Pedersen MS et al (2008) Convolutive blind source separation methods. Springer handbook of speech processing. Springer, Berlin, Heidelberg, pp 1065–1094

14. Blouet R, Rapaport G, Févotte C (2008) Evaluation of several strategies for single sensor speech/music separation. In: Proceedings of IEEE international conference on acoustics, speech and signal processing—proceedings, pp 37–40. https://doi.org/10.1109/ICASSP.2008.4517540
15. Ozerov A, Févotte C (2009) Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. In: Proceedings of IEEE transactions on audio, speech, and language processing, vol 18, no 3
16. Grais EM, Erdogan H (2011) Single channel speech music separation using nonnegative matrix factorization and spectral masks. In: Proceedings of 17th international conference on digital signal processing (DSP). IEEE (2011)
17. Virtanen T (2007) Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. Proc IEEE Trans Audio, Speech Lang Proc 15(3):1066–1074
18. Schmidt MN, Mørup M (2006) Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In: Proceedings of international conference on independent component analysis and signal separation. Springer, Berlin, Heidelberg (2006)
19. Lefevre A, Bach F, Févotte C (2012) Semi-supervised {NMF} with time-frequency annotations for single-channel source separation. In: Proceedings of 13th international society for music information retrieval
20. Huang P-S, Kim M, Hasegawa-Johnson M, Smaragdis P (2015) Joint optimization of masks and deep recurrent neural networks for monaural source separation. In: Audio speech language processing, in proceedings of IEEE/ACM Transactions on. 23https://doi.org/10.1109/TASLP.2015.2468583
21. Fukayama Y (2014) A modified Wiener filter suitable for separation of individual instrumental sounds in monaural music signals. In: Proceedings of IEEE conference on norbert wiener in the 21st Century (21CW), Boston, MA, 2014, pp. 1–4. https://doi.org/10.1109/NORBERT.2014.6893934
22. Fukayama Y (2019) Separation of individual instrumental tones in monaural music signals applying a modified wiener filter and the gabor wavelet transform. In: Proceedings of the ISCIE international symposium on stochastic systems theory and its applications, vol 2019. The ISCIE symposium on stochastic systems theory and its applications
23. Le Roux J, Vincent E (2012) Consistent wiener filtering for audio source separation. Proc IEEE Sig Proc Lett 20(3):217–220
24. Gunawan D, Sen D (2009) Music source separation synthesis using multiple input spectrogram inversion. In: Proceedings of IEEE international workshop on multimedia signal processing
25. Bofill P, Zibulevsky M (2000) Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform. In: Proceedings of ICA, vol 2000
26. Bofill P, Zibulevsky M (2001) Underdetermined blind source separation using sparse representations. Proc Sig Process 81(11):2353–2362
27. Huang, P-S et al (2012) Singing-voice separation from monaural recordings using robust principal component analysis. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP)
28. Ikemiya Y, Yoshii K, Itoyama K (2015) Singing voice analysis and editing based on mutually dependent F0 estimation and source separation. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE
29. Liutkus A et al (2012) Adaptive filtering for music/voice separation exploiting the repeating musical structure. In: Proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE
30. Rafii Z, Pardo B (2012) Music/voice separation using the similarity matrix. In: Proceedings of ISMIR
31. Rafii Z, Pardo B (2012) Repeating pattern extraction technique (REPET): A simple method for music/voice separation. Proc IEEE Trans Audio Speech Lang Process 21(1):73–84
32. Lin Y-P et al (2010) EEG-based emotion recognition in music listening. Proc IEEE Trans Biomed Eng 57(7):1798–1806

33. Markov K, Matsui T (2014) Music genre and emotion recognition using Gaussian processes. In: Proceedings of IEEE access 2
34. Fukayama S, Goto M (2016) Music emotion recognition with adaptive aggregation of Gaussian process regressors. In: Proceedings of international conference on acoustics, speech and signal processing (ICASSP). IEEE
35. Chordia P, Rae A (2007) Raag recognition using pitch-class and pitch-class dyad distributions. In: Proceedings of ISMIR
36. Sridhar R, Geetha TV (2009) Raga identification of carnatic music for music information retrieval. Proc Int J Recent Trends Eng
37. Dighe P, Karnick H, Raj B (2013) Swara histogram based structural analysis and identification of Indian classical ragas. In: Proceedings of ISMIR
38. Padmasundari G, Murthy HA (2017) Raga identification using locality sensitive hashing. In: Proceedings of twenty-third national conference on communications (NCC). IEEE
39. Dutta S, Krishnaraj Sekhar PV, Murthy HA (2015) Raga verification in carnatic music using longest common segment set. In: Proceedings of ISMIR (2015)
40. Katte (2013) Multiple techniques for raga 378 identification in Indian classical music. Int J Electron Comput Eng (2013)
41. Gulati S et al (2016) Phrase-based rāga recognition using vector space modeling. In: 2016 In proceedings of IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE
42. Kan A, sankar AS, Sundhar S (2017) A comparison of machine learning approaches to classify tala COMP 562
43. Koduri GK et al (2014) Intonation analysis of rāgas in Carnatic music. J New Mus Res
44. Koduri GK, Joan SJ, Serra X (2012) Characterization of intonation in carnatic music by parametrizing pitch histograms. In: Gouyon F, Herrera P, Martins LG, Müller M. ISMIR 2012: proceedings of the 13th international society for music information retrieval conference; 2012 Oct 8–12; Porto, Portugal. Porto: FEUP Ediçoes, 2012. International society for music information retrieval (ISMIR) (2012)
45. Subramanian M (2007) Carnatic ragam thodi-pitch analysis of notes and gamakams. J Sangeet Natak Akademi 41(1)
46. Guedes C, Trochidis K, Anantapadmanabhan A (2018) Modeling carnatic rhythm generation: a data driven approach based on rhythmic analysis. In: Proceedings of the 15th sound & music computing conference
47. Ganguli K, Guedes, C (2019) An approach to adding knowledge constraints to a data-driven generative model for Carnatic rhythm sequence. Trends Electr Eng 9(3)
48. Heshi R et al (2016) Rhythm and timbre analysis for carnatic music processing. In: Proceedings of 3rd international conference on advanced computing, networking and informatics. Springer, New Delhi
49. Trochidis K et al (2016) CAMeL: carnatic percussion music generation using n-gram models. In: Proceedings of 13th sound and music computing conference (SMC), Hamburg, Germany, vol 31
50. Sebastian J, Murthy HA (2017) Onset detection in composition items of carnatic music. In: Proceedings of international society for music information retrieval ISMIR
51. Bellur A, Murthy HA (2013) A novel application of group delay function for identifying tonic in Carnatic music. In: Proceedings of 21st european signal processing conference (EUSIPCO 2013). IEEE (2013)
52. Samsekai Manjabhat S, et al (2017) Raga and tonic identification in carnatic music. J New Music Res 46(3)
53. Salamon J, Gulati S, Serra X (2012) A multipitch approach to tonic identification in indian classical music. In: Gouyon F, Herrera P, Martins LG, Müller M (eds) ISMIR 2012: Proceedings of the 13th international society for music information retrieval conference
54. Bellur A et al (2012) A knowledge based signal processing approach to tonic identification in Indian classical music. In: Serra X, Rao P, Murthy H, Bozkurt B (eds) Proceedings of the 2nd CompMusic Workshop; 2012 Jul 12–13; Istanbul, Turkey. Barcelona

55. Mammen S et al (2016) iSargam: music notation representation for Indian Carnatic music. EURASIP J Audio Speech Music Proc 2016(1)
56. Bellini X, Nesi P (2001) Wedel music format: an XML music notation format for emerging applications. In: Proceedings of first international conference on WEB delivering of music, IEEE computer society, Washington
57. Hoos HH, Hamel KA, Renz K, Kilian J (2001) Representing score-level music using the GUIDO music notation format. Computing in Musicology, MIT Press, Cambridge, p 12
58. Sridhar R, Geetha TV (2006) Swara identification for South Indian classical music. In: Proceedings of 9th international conference on information technology (ICIT'06), Bhubaneswar, pp 143–144
59. Prashanth TR, Venugopalan R (2011) Note identification in Carnatic Music from Frequency Spectrum In: Proceedings of international conference on communications and signal processing, Calicut, pp 87–91
60. Ranjani HG, Arthi S, Sreenivas TV (2011) Carnatic music analysis: Shadja, swara identification and rAga verification in AlApana using stochastic models. In: Proceedings of IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), New Paltz, NY, pp 29–32
61. Bunt L, Stige B (2014) Music therapy: an art beyond words. Routledge
62. Maratos A et al (2008) Music therapy for depression. Cochrane database of systematic reviews, vol 1
63. Warwick A, Alvin J (1991) Music therapy for the autistic child. Oxford University Press (1991)

Effects of Vocal Loading on Singing Power Ratio and Singer's Formant with Indian Heavy Metal Vocalists

Mahesh Soumya and S. V. Narasimhan

Abstract Specialized voice techniques, conventionally referred to as "growling" or "screaming", are frequently used by heavy metal singers. These techniques are used to enhance their expressivity and project their voices. These singers use various adjustment of vocal fold to yield harmonically more vibrant voices. However, most of the heavy metal vocalists tend to overuse their vocal systems, which can have an adverse effect on their quality of voice. The objective of this study was to document the impact of vocal loading on singing power ratio and singer's formant with Indian heavy metal vocalists. Participants were instructed to phonate /a/ at their comfortable pitch, and the loudness and the voice samples were recorded in two conditions: before (pre-fatigue condition) and after singing performance (post-fatigue condition). These samples were analysed acoustically, and the singing power ratio and the singer's formant were extracted and compared statistically. Results of the study revealed a higher singing power ratio for the pre-fatigue condition compared to the post-fatigue condition and the singer's formant did not show significant differences between the pre-fatigue and post-fatigue condition.

Keywords Singing power ratio · Singer's formant · Heavy metal singers

1 Introduction

Singers voice can be considered as the most proficient musical instrument of all that enhances the verbal expression of spoken language [1]. The requisition of accurate control of pitch during singing needs profound knowledge and attention towards behaviours like resonance training, which is frequently witnessed in singing but rare in speech.

Singing is a sensory-motor phenomenon entailing a precise balance between various physical skills, such as resonance, phonation and respiration [2]. Excellent coordination between all the subsystems of speech is a primary requisite for singing and changes for every musical tone and various singing styles. A singer's voice can

M. Soumya (✉) · S. V. Narasimhan
JSS Institute of Speech & Hearing, Mysore, Karnataka, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_2

produce an extensive range of sounds, and this range is replicated in the various styles of singing. Maintaining the rhythm and the melodic aspects, maintaining an appropriate tone and effective management of breath are a part of excellent vocal performance or singing. This accurate balance between skills frequently develops as a product of experience and training, thereby establishing alterations in the singer's voice in terms of their acoustic, physiological and perceptual aspects.

Various studies have documented the occurrence of a singer's formant in the voice of professionals [3]. According to Koichi Omori et al., professional singers exhibit a "ringing" quality of voice that corresponds to the singer's formant in the spectrum especially during the singing performance. This "ringing" quality allows the singers to project the voice, without any amplification [4]. This is due to the way the sound energy of a trained singer's voice is distributed across various frequencies. The singer's formant is a perceptual correlate of "ringing" voice quality ("vocal ring"), "twang" and "resonant voice". It is a peak in the spectral envelope around 3 kHz [5] but it has also been identified as a collection of the fifth, fourth and third formant frequencies in the spectrum, rather than a single "extra" formant [6]. Singing power ratio (SPR) is one of the measures to quantify the singer's formant. It is the difference between the amplitude of the most significant peak in the spectrum between 2 and 4 kHz to that between 0 and 2 kHz [3]. SPR offers a quantifiable measurement for assessing the voice quality in singers, and it demonstrates a typical relationship with the duration of voice training [4].

Many studies have attempted to quantify "ring" in the singing voice and have reported that SPR was higher with untrained talented singers compared to untrained non-talented singers [7]. Studies have also reported a similar SPR between professional and non-professional singers [4]. In the existing literature, the singer's formant and SPR in singers have been documented primarily from the western classical tradition or Indian Carnatic or Hindustani music singers. Apart from these groups, an attempt to investigate vocal aspects in various genres of singers is still limited. Only a handful of studies has been published on contemporary "commercial" music, such as heavy metal.

The heavy metal genre is a musical style that came into existence around the 1970s. Heavy metal vocalists usually overdrive their voices, tend to sing long sustained notes and produce a heavy vibrato for projecting their voice. Heavy metal singing includes various laryngeal alterations, of which the main one is "grunting". Periodic vibrations may often be found in conventional singing, and aperiodic vibrations such as distortion can be evidenced in the heavy metal genre. Distortions are frequently used to add an expressivity tinge and as a timbral ornament to the singing voice. Distortion is blended with different vocal registers in heavy metal singing [8]. Specialized voice techniques conventionally referred to as "growling" or "screaming" are commonly used by singers in especially hard rock or heavy metal music genres. These techniques do not only add distinctive voice quality but also help in projecting the singer's voice [9].

Metal singing also involves modulation of the airflow passing through the ventricular folds, which could either be produced through inhalation or exhalation [10]. The

different directions of airflow can be thought of as parallel to the linguistic distinction made between voiced and unvoiced methods of articulating consonants. When performing exhaled vocals, one's larynx vibrates, indicating that the vocal cords are vibrating rather forcefully, whereas this vibration does not occur with inhaled vocals. This difference is said to have an intense effect on the overall sound quality produced, the effort with which different phonetic articulations can be uttered, the ability of a vocalist to sustain notes using screaming and the degree of strain on the voice. Wei-Ho Tsai et al. perceptually evaluated the growl-like voices of Western metal singers and noticed predominantly harsh, whispery voice throughout the song [11]. These professional singers, in addition to using muscular tension adjustments, demonstrate higher levels of laryngeal tension. Nevertheless, the quality of this professional breathy voice is often supplemented by a stronger energy level. The additional force which is produced by exhaled screams is discernible particularly exhausting for the performer and more destructive than inhaled screams. Considering all these adverse effects on the vocal behaviour of the metal singers, the spectral and acoustical measures of the vocal attributes are bound to get affected.

Thus, it can be inferred that most of the heavy metal vocalists tend to overuse or use their vocal systems differently for singing, and this could have an adverse effect on their voice quality [12]. With the increasing popularity of the metal singing genre and a growing number of metals singers in India [13], there is a need to investigate the singer's formant and SPR in the voice of Indian heavy metal singers. Furthermore, the studies that have documented the effect of the vocal loading or vocal fatigue on the SPR or the singer's formant, especially among Indian heavy metal singers, are limited in availability. Therefore, the current study is aimed at documenting the effect of vocal loading on SPR and singers' formants with Indian heavy metal vocalists.

2 Method

2.1 Subjects

The study included ten male heavy metal vocalists (extreme metal screamers) of Indian origin, who were professionally trained for at least five years. The number of participants was determined based on the statistical power analyses using G-Power (ver. 3) [14]. The power of the test was assigned a value of 0.8 and the α error probability of 0.05. All the subjects were in the age range of 25–32 years. A questionnaire was administered in English to obtain the details on their phonatory habits, work habits, previous history of voice problems, tobacco usage, voice therapy, or surgical or medical intervention to all the subjects. The responses from the questionnaire were examined and the subjects having any recent history or current complaints of respiratory tract infections, laryngeal surgeries, tobacco usage, hearing loss or family history of voice disorders were not involved in the present study. Informed consent

indicating the willingness to participate in the study was obtained from the parents of each participant.

2.2 Procedure

A sound-treated room used for audiometry which met the ANSI specifications prescribed for hearing evaluation [15] was used for data collection. The room was free from distraction, had a comfortable temperature, and was well lit. The subjects were seated comfortably in front of the laptop screen during the recording. Participants were instructed to phonate the vowel /a/three times at their habitual pitch and comfortable loudness levels. Data was collected in two separate instances: condition 1 was thirty minutes before the jam session or the vocal performance started, and condition 2 was within thirty minutes after the completion of the jam session or vocal performance. A dynamic, unidirectional microphone with a frequency response of 60–12,000 Hz was used to record all the voice samples. The participant's mouth to microphone distance was 10 cm and the voice was recorded on the Praat software (Version 5.3.23).

2.3 Analysis

A steady portion of 6 to 8 ms of each vowel /a/ was selected to analyse the SPR and singer's formant in the phonation samples. IMpx software was used to analyse the phonation samples for spectral energy distribution. Harmonics peaks were identified in the spectrum envelope between 0–2 and 2–4 kHz. The amplitude of the highest amplitude obtained in the region of lower harmonics (0–2 kHz), the amplitude of the highest amplitude obtained in the region of higher harmonics (2–4 kHz) and peak frequency of the highest amplitude obtained in the region of higher harmonics (2–4 kHz) was extracted. The power ratio of the most significant peak between higher harmonics (2–4 kHz) and the lower harmonic (0–2 kHz) region was calculated and expressed in decibels as SPR.

3 Results and Discussion

The tabulated results were subjected to both descriptive and inferential statistics. As a part of descriptive statistics, the mean and standard deviation (SD) of the singer's formant and SPR were carried out for both condition 1 and condition 2. The results of the descriptive statistics are displayed in Table 1 (Fig. 1).

The results of the descriptive statistics revealed that the mean scores of SPR for condition 1 were found to be higher when compared to condition 2; however, the

Table 1 Singer's formant and SPR for condition 1 and condition 2 with Indian heavy metal vocalists

S. No.	Parameter	Condition 1		Condition 2	
		Mean	SD	Mean	SD
1	SPR (in dB)	−1.71	1.10	−5.90	5.76
2	Singer's formant (in Hz)	3379	250.92	3619	320.96

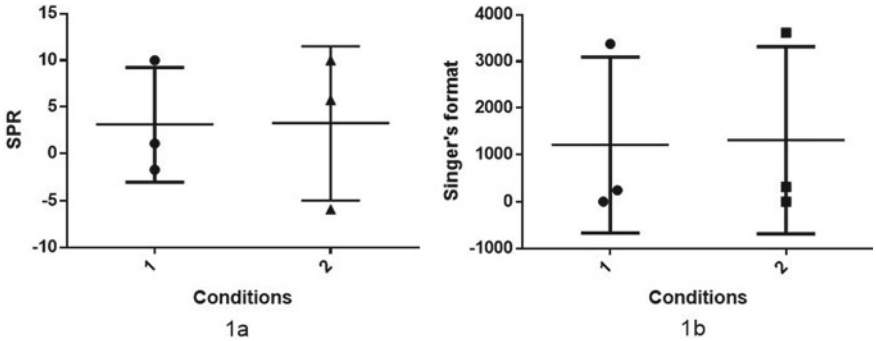


Fig. 1 Mean and standard deviation of SPR (1a) and singer's formant (1b) in both conditions

ratios obtained for both the conditions were negative. Contrastingly, mean singer's formant was higher in condition 2 compared to condition 1. Standard deviation values of both SPR and singer's formant were higher in condition 2 compared to condition 1, showing higher variability of SPR and singer's formant in condition 2 compared to condition 1. Furthermore, the paired sample t-test was carried out as a part of inferential statistics to compare SPR and singer's formant between both the conditions. The results of the inferential statistics showed that there was a significant difference in SPR ($/t/ = 2.462$, $/p/ = 0.032$) between the two conditions; however, the singer's formant did not show any significant difference. Thus, it was inferred from the statistical analysis that the SPR in condition 1 was significantly higher compared to condition 2 and there was no significant difference in singer's formant between both the conditions.

The results of the present study indicated that the SPR reduced with Indian heavy metal singers after the performance or vocal loading. Earlier investigations have also reported a higher SPR as a result of higher energy within the 2–4 kHz region and a perceptual reflection of a good singing voice quality, whereas a lower energy relationship between the higher and lower harmonics resulted in lower SPR and reflected poor singing voice quality [4, 7, 16].This can be due to varied physiological competency while singing, especially with heavy metal singers.

Heavy metal is an aggressive subgenre of metal music. The singers naturally avoid pitched singing in support of specialized screaming techniques such as growls, grunts, creak and distortion. Ornamentations, vibratos and breath sounds can also be

seen as effects. These extreme vocal effects are produced by the use of the ventricular folds, to vary the length of the vocal tract, along with exhalation phonation [17]. Metal singing also involves modulation of the airflow passing through the ventricular folds which either could be produced through inhalation or exhalation [10]. Perceptual evaluation of growl-like voices of Western metal singers showed that the voice of the metal singers tends to be predominantly harsh and whispery throughout the song [11]. These professional singers, in addition to using muscular tension adjustments, demonstrate higher levels of laryngeal tension. The additional force which is produced by exhaled screams is discerned to be particularly exhausting for the performer and more destructive than inhaled screams. Considering all these adverse effects of the vocal behaviour of metal singers, the spectral measures such as SPR are bound to get affected.

The results of this study also outlined the disparity within the singer's formant in pre- and post-fatiguing condition. Although there was no significant difference noted in the singer's formant between the pre- and post-fatiguing conditions, higher formant energy was observed in the post-fatigue condition compared to the pre-fatigue condition. The present study also evidenced the singer's formant ranging from 3 to 4 kHz. These findings are consistent with previous results reporting centre frequency of the singer's formant ranging from 2.2 to 3.8 kHz [5, 18, 19]. Thus, it can be inferred that the power of the most significant harmonics peak in the frequency range between 3 and 4 kHz is vital for singing voice quality among Indian heavy metal singers.

4 Conclusion

The objective of the present study was to document the effect of vocal loading on SPR and singer's formant with Indian heavy metal vocalists. The present study involved the calculation of SPR and the singer's formant in two conditions. Voice samples were recorded before (pre-fatigue condition) and after (post-fatigue condition) the singing performance. Furthermore, the results of the current study showed high SPR for the pre-fatigue condition compared to post-fatigue condition and no significant change in the singer's formant frequency in both conditions. Thus, SPR can be considered as a quantitative measurement for estimating singing voice quality as it provides a distinctive correlation with the effect of vocal loading on Indian metal singers. Further studies can shed more light on calculating these spectral measures with heavy metal singers during their vocal performance, as this singing style involves specialized screaming techniques, such as growls, grunts, squeals, as well as distortion when compared to non-singers.

References

1. Kob M, Henrich N, Herzel H, Howard D, Tokuda I, Wolfe J (2011) Analysing and understanding the singing voice: recent progress and open questions. Curr Bioinform 6:362–374
2. Bunch M (1982) Dynamics of singing voice. NY
3. Watts CR, Murphy J, Barnes-Burroughs K (2003) Pitch matching accuracy of trained singers, untrained subjects with talented singing voices, and untrained subjects with nontalented singing voices in conditions of varying feedback. J Voice 17:185–194
4. Omori K, Kacker A, Carroll LM, Riley WD, Blaugrund SM (1996) Singing power ratio: quantitative evaluation of singing voice quality. J Voice 10:228–235
5. Sundberg J (2015) Round table on nasal resonance in singing. In: Manfredi C (ed) Pan European voice conference. Firenze University Press, Firenze, p 154
6. Cleveland TF (1994) A clearer view of singing voice production: 25 years of progress. J Voice 8:18–23. https://doi.org/10.1016/S0892-1997(05)80315-7
7. Watts C, Barnes-Burroughs K, Estis J, Blanton D (2006) The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers. J Voice 20:82–88
8. Meireles AR (2016) Perceptual and acoustic study of voice quality in high-pitched heavy metal singing. In: Proceedings of the international conference on speech prosody, pp 1245–1249
9. Nieto O (2013) Unsupervised clustering of extreme vocal effects. In: Proceedings of advances in quantitative laryngologyconference, pp 115–116
10. Severin F, Bozkurt B, Dutoit T (2005) HNR extraction in voiced speech, oriented towards voice quality analysis. In: 2005 13th European signal processing conference, pp 1–4
11. Tsai WH, Rodgers D, Wang HM (2004) Blind clustering of popular music recordings based on singer voice characteristics. Comput Music J 28:68–78
12. Walser R (1993) Running with the Devil: power, gender, and madness in heavy metal music. Wesleyan University Press
13. Eckers C, Hütz D, Kob M, Murphy P, Houben D, Lehnert B (2009) Voice production in death metal singers. Nag/Daga 1747–1750
14. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191
15. ANSI S3.1 (2013) Maximum permissible ambient noise levels for audiometric test rooms. The American National Standards Institute
16. Lundy DS, Roy S, Casiano RR, Xue JW, Evans J (2000) Acoustic analysis of the singing and speaking voice in singing students. J Voice 14:490–493. https://doi.org/10.1016/S0892-1997(00)80006-5
17. Smialek E, Depalle P, Brackett D (2012) Musical aspects of vowel formants in the extreme metal voice. Dafx12.York.Ac.Uk. 1–8
18. Dmitriev L, Kiselev A (1979) Relationship between the formant structure of different types of singing voices and the dimensions of supraglottic cavities. Folia Phoniatr Logop 31:238–241
19. Seidner W, Schutte H, Wendler J, Rauhut A (1983) Dependence of the high singing formant on pitch and vowel in different voice types. In: Proceedings of the Stockholm music acoustics conference, pp 261–268

Audio Quality Control Method Based on ASS (Audio Secret Sharing)

Sachin Kumar Singh and Mainejar Yadav

Abstract Audio Secret Sharing (ASS) can be used to provide security to audio channels by dividing the audio into smaller units called shares. With the advancement of technology and the availability of ubiquitous devices, a lot of audio platforms came into the picture that provides speeches, music, and other audio files which are usually guarded by a paywall. In cases where the low quality of the audio is provided for free and high-quality audio is guarded by a paywall, in that scenario the vendor stores two separate copies of the audio, one low and one high which certainly acquire twice space. The proposed scheme offers to store low-quality audio which can generate high quality with help of a share, this comes out to be space-efficient and we also don't need to maintain two separate copies of the same audio. In the proposed scheme, an audio quality control method has been defined that can be used in modern audio platforms.

1 Introduction

Technology is advancing at an enormous pace, and there are unprecedented platforms on the rise. One major of them is the audio platform that provides audio such as music, audiobooks, and speeches to the end-user. They may or may not offer a small version or partial audio for free as a trial version, and then they possess a pay-wall for the complete version they usually don't have any quality control system.

Audio Secret Sharing (ASS) can be seen as an extension of *Visual Secret Sharing (VSS)* [12] proposed by Shamir et al. [1] and Blakey et al. [3] in 1979. The audio consists of a stream of amplitudes, and we will be applying *(2, 2)* secret sharing to an audio file. That will generate two shares one will be low-quality audio and other share known as the quality share will act as a key for enhancing the audio quality. This scheme not limited to *(2, 2)* it can be further extended to *(n, n)*. If some intruder

S. K. Singh (✉) · M. Yadav
Department of Computer Science and Engineering, Rajkiya Engineering College, Sonbhadra, Uttar Pradesh, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_3

tries to enhance the quality of the low-quality share without the quality share, it will even further degrade the low-quality share.

The remaining part of the paper is as follows: Literature Survey is presented in Sect. 2. The Proposed Method is described in Sect. 3, Analysis of Experiment result is performed in Sect. 4, and the paper is concluded in Sect. 5.

2 Literature Review

In this section, a review of existing Audio Secret Sharing (ASS) is presented.

The early *Audio Secret Sharing (ASS)* schemes were based on the interference property of the wave. The audio file was considered as a wave and binary secret was embedded in the wave of the audio. For embedding the binary secret audio was used as a cover, such *ASS* schemes were proposed by Desmedt et al. [4], Yang [6], and Lin et al. [5]. The major disadvantage with these schemes was that they were only applicable to the binary secret which eventually provides very low efficiency in terms of space utilization. Later on, the audio was used as a secret in a *(2, n) ASS* scheme was proposed by Ehdaie et al. [7]. Instead of wave nature of audio in this scheme amplitude properties of audio have been exploited. In this scheme a matrix is randomly generated and which leads to share generation. But if the matrix which is randomly generated possesses a singular sub-matrix then matrix needs to be regenerated. A method based on ANDing and ORing was proposed by Prabir et al. [13]. For share generation and reconstruction, ANDing and ORing were used. This offers the benefit of less computational complexity but a cover is required and it cannot be used for real-time communications. Bharti et al. [11] proposed a method for ASS which can be applicable to real-time audio systems which was based on dividing the audio stream of amplitude in equal shares. Due to the use of the scaling factor is this approach they correlation coefficient don't remain intact and there is a minor quality loss, which can be significant is the adjacent amplitude in the audio stream are close. Singh et al. [16] proposed a *(k, n)* ASS with a share verification technique, it also doesn't possess any quality control method.

Apart from the above-mentioned limitations, no authors have provided a method for quality control of audio. In this paper, the quality control method has been proposed and other limitations found in the literature are also reduced.

3 Proposed Method

In the context of this work low quality share is the degraded original audio and quality share is a share (consisting amplitudes) which is combined in a specific manner with the low quality share for regenerating original audio.

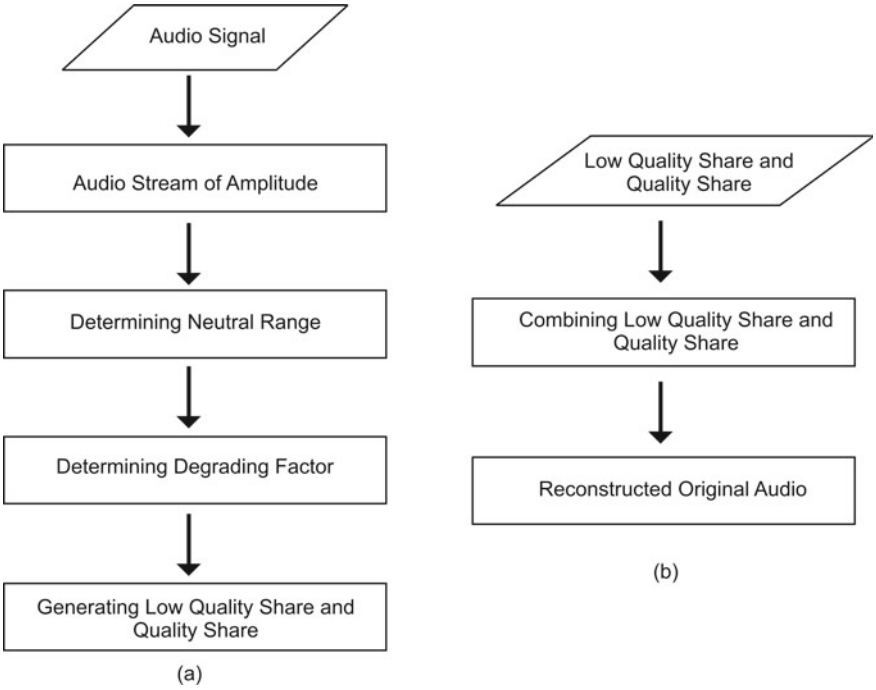


Fig. 1 **a** Flowchart for generating Low Quality Share and Quality Share, **b** Flowchart for reconstruction of original audio

For generating the low-quality share and quality share from the original audio signal we need to first determine the Degrading factor, Neutral value, and Neutral range. The degrading factor is a measure of the extent to which we want to degrade the audio quality from the audio. It is a range of factors to pick from. This range has been used to provide randomness to the share generation and making it more robust. This range should be picked with great precision. If the values are too low then there will not be any significant change in the quality of the audio, if the values are too high then it will make the audio not audible and only consist of noise (Fig. 1).

While degrading the audio signal it should be kept in consideration that silence should remain silent, which means that the zero and close to zero amplitude should remain the same, this value is known as neutral value. Neutral value depends on the representation used by the system that is the available range of amplitude. For ex-ample, if the system is using int8 or int16 representation the neutral point will be zero. If system uses uint8 (ranging 0 to 255) representation then neutral point will be 128. The neutral range is a small range in which we will not perform any changes to the amplitude of the audio, if we do so it will lead to unnecessary background noise.

Algorithm 1 Construction of Low-Quality Share (LQ_{Share}) and Quality Share(Q_{Share})

Input : Original Audio Stream of Amplitudes (O^A) of size $NX1$, Range of *Degrading factor* (V_{Factor}), *Neutral Range* (here 110 to 140)
Output : Low-Quality Share (LQ_{Share}) and Quality Share (Q_{Share})

```
for j ← 1 : N do
    x = random(V_Factor)
    x = random(0,1)
    if sign = 1 then
        if i>110 and i <140 then
            LQ_Share[i] = O_A[i]                          /*In Neutral Range*/
            Q_Share[i] = 0
        else
            LQ_Share[i] = O_A[i] + ((O_A[i])*x)
            Q_Share[i] = LQ_Share[i] - O_A[i]
        end if
    else
        if i>110 and i <140 then
            LQ_Share[i] = O_A[i]                          /*In Neutral Range*/
            Q_Share[i] = 0
        else
            LQ_Share[i] = O_A[i] - ((O_A[i])*x)
            Q_Share[i] = O_A[i] - LQ_Share[i]
        end if
    end if
end for
```

Algorithm 2 Reconstruction of audio with Low-Quality Share (LQ_{Share}) and Quality Share(Q_{Share})

Input : Low-Quality Share (LQ_{Share}) and Quality Share (Q_{Share})
Output : *Reconstructed Audio*(R_A)

```
N = size(LQ_Share)
for j ← 1 : N do
    R_A[i] = LQ_Share[i] + Q_Share[i]
end for
```

For each amplitude in the audio signal, a random degrading factor is picked from the available range. Then the product of amplitude and picked degrading factor is either subtracted or added to the original amplitude (sign bit is used for determining that subtraction will be performed or addition will be performed). Now, this value is stored in the Low-Quality Share and the difference in the value of amplitude in the quality share, and the original amplitude is stored in the Quality Share. If the amplitude lies within the neutral range then no such operations are performed and the original amplitude value is stored in the Low-Quality Share and zero is stored in Quality Share. Experiment it has been determined Degrading Factor (V_{Factor}) should

Table 1 Generation of low-quality share and quality share

Amplitude Stream (O_A)	145	32	56	69	201
Degrading Factor(V_{Factor})	0.2	041	0.1	0.05	0.8
In Neutral Range	Yes	No	No	No	No
Low Quality Share(LQ_{Share})	145	19	50	66	40
Quality Share (Q_{Share})	0	13	6	3	161

range around 0.05 to 1 for mild quality degradation. It could be further adjusted for low or strong quality degradation.

Let's consider an audio stream consisting of amplitude $O_A = [145, 32, 56, 69, 201]$, neutral range = [110–140], Degrading factor (V_{Factor}) range = [0.05 to 1] (Table 1). So for each amplitude in the audio stream, a degrading factor is picked form the range and product of amplitude, and picked degrading factor is either subtracted or added to the original amplitude depending on the nature of the sign bit (1 for addition and 0 for subtraction). If the amplitude lies in the neutral range then no changes will be made. So, for the framed example Low-Quality Share and Quality Share can be generated.

The quality of the Low-Quality Share will be less than the original audio which will be later determined by MOS and the correlation coefficient between the original audio and Low-Quality Share will not be equal to 1, which indicates loss of content as compared to the original audio.

4 Analysis of Experiment Result

4.1 Database Used for the Experiment

IndicTTS [2] database has been used. The audio is taken from the database are of various durations. IndicTTS database contain male and female pronounced sentences possess in many different Indian languages.

4.2 Performance Measurements

Comparative analysis has been performed on the original and low-quality share audio. For performing this comparative analysis subjective as well as objective techniques have been used. Comparison Mean Opinion Score (CMOS), Correlation Coefficient and PESQ (Perceptual Evaluation of Speech Quality) have been used. For calculating CMOS which is a subjective parameter, 15 persons are considered divided into three equal groups of five person each. Audios has been played in each group and rating has

Table 2 The rating scale used for CMOS [14, 15]

Rating	Label
3	Very Good
2	Good
1	Slightly Good
0	About the same
−1	Slightly Worse
−2	Worse
−3	Much Worse

been received from each individual. Persons ratings will be based on the comparison, they will compare the audio quality and rate according to the labels in table 2. Ratings are taken twice, firstly by playing Original audio and low-quality share audio and secondly by playing Original audio and reconstructed audio (audio generated by combining low-quality share and quality share). Rating received will vary from −3 to 3. Finally, the average rating is calculated. The correlation coefficient establishes a correlation between the audios. It varies from 1.0 to −1.0. PESQ—Perceptual Evaluation of Speech Quality (Raw MOS, MOS-LQO) which is an ITU standard is also been evaluated for both the original and reconstructed audio. Raw MOS range between −0.5 to 4.5 and MOS-LOQ range between 1 and 5.

The experiment is conducted in five cases. These cases are based on real scenarios.

1. Case: 1 No distortion is performed during transmission or at the user end.
2. Case: 2 Intruder attempt to enhance quality by enhancing amplitude by fixed or random value.
3. Case: 3 Intruder attempts to enhance quality by diminishing amplitude by a fixed or random value.
4. Case: 4 Intruder attempts to enhance the quality by making amplitude A_m, cA_m where c is a constant.
5. Case: 5 Intruder attempts to enhance quality by enhancing or diminishing amplitude by a fixed or random value.

4.3 Experiment Result

On an input audio (Fig. 2a), algorithm 1 is applied and low-quality and quality shares are generated shown in Fig. 2b. Reconstruction of original audio is performed by applying algorithm 2 on low-quality share and quality share, reconstructed audio is shown in Fig. 2c Case 1, no distortion is performed. For the input audio signal shown in Fig. 2, two shares have been generated by using Algorithm 1. One is the quality share (which will act as a key for quality enhancement) and the other is low-quality share (which is the low-quality audio) shown in Fig. 3. In this case, no modification

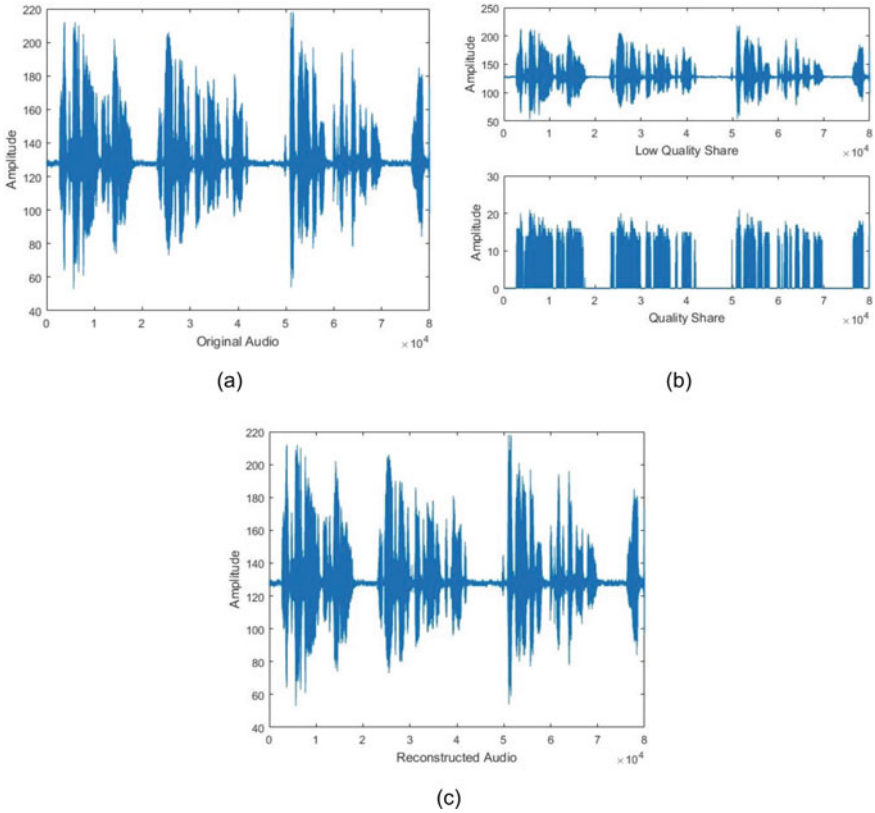


Fig. 2 **a** Input Audio, **b** Low Quality Share (LQ_{Share}) and Quality Share (Q_{Share}), **c** Reconstruction Audio

Fig. 3 Reconstruction audio

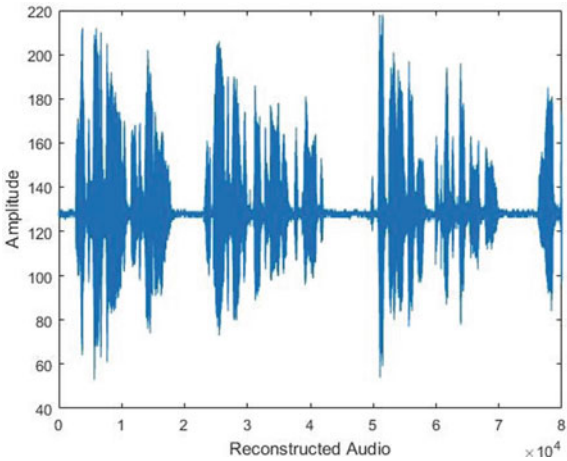


Table 3 CMOS rating for original and low-quality share audio

Group	Person 1	Person 2	Person 3	Person 4	Person 5
Group 1	−1	−2	−1	−1	−2
Group 2	−1	−1	−1	−1	−1
Group 3	−1	−2	−1	−2	−1

Table 4 CMOS rating for original and reconstructed audio

Group	Person 1	Person 2	Person 3	Person 4	Person 5
Group 1	0	0	−1	1	0
Group 2	0	0	0	0	0
Group 3	0	0	0	0	0

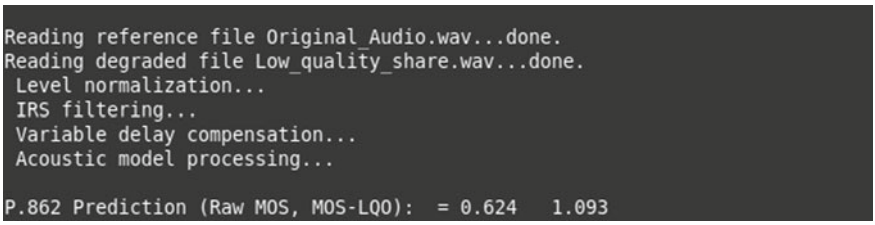


Fig. 4 PESQ (Raw MOS, MOS-LQO) for original and low-quality share audio

is assumed during transmission or at the user end. Reconstruction of the original audio is performed by algorithm 2. Reconstructed audio has been shown in Fig. 2.

CMOS rating has been shown in Table 3 for original and low-quality share audio and Table 4 for original audio and reconstructed audio (low-quality + quality share). For low-quality and share audio average CMOS comes out to be −1.266, according to the label which indicates "Slightly worse" which shows degradation in audio quality and for original audio and reconstructed audio comes out to be 0 (Zero), according to the label which indicates "About the same" which shows no degradation in audio quality. The correlation coefficient between low quality and original audio found to be 0.9998 and for original audio and reconstructed audio remains intact and equals to 1.

PESQ (Raw MOS, MOS-LQO) comes to 0.624 and 1.093 for original and low-quality share audio (Fig. 4), which indicates degraded quality of audio and it comes be to 4.500 and 4.549 for original audio and reconstructed audio (Fig. 5) that shows the reconstructed audio quality didn't degrade and possess the same quality as original audio.

Case 2, intruder attempts to enhance quality by increasing amplitude by a fixed or random value. Figure 6a shows the low-quality share audio after intruders attempt to enhance. The correlation coefficient comes out to be 0.9993 which indicates more degradation in quality than Low-Quality Share.

```
Reading reference file Original_Audio.wav...done.
Reading degraded file Reconstructed_Audio.wav...done.
 Level normalization...
 IRS filtering...
 Variable delay compensation...
 Acoustic model processing...

P.862 Prediction (Raw MOS, MOS-LQO):  = 4.500   4.549
```

Fig. 5 PESQ (Raw MOS, MOS-LQO) for original and reconstructed audio

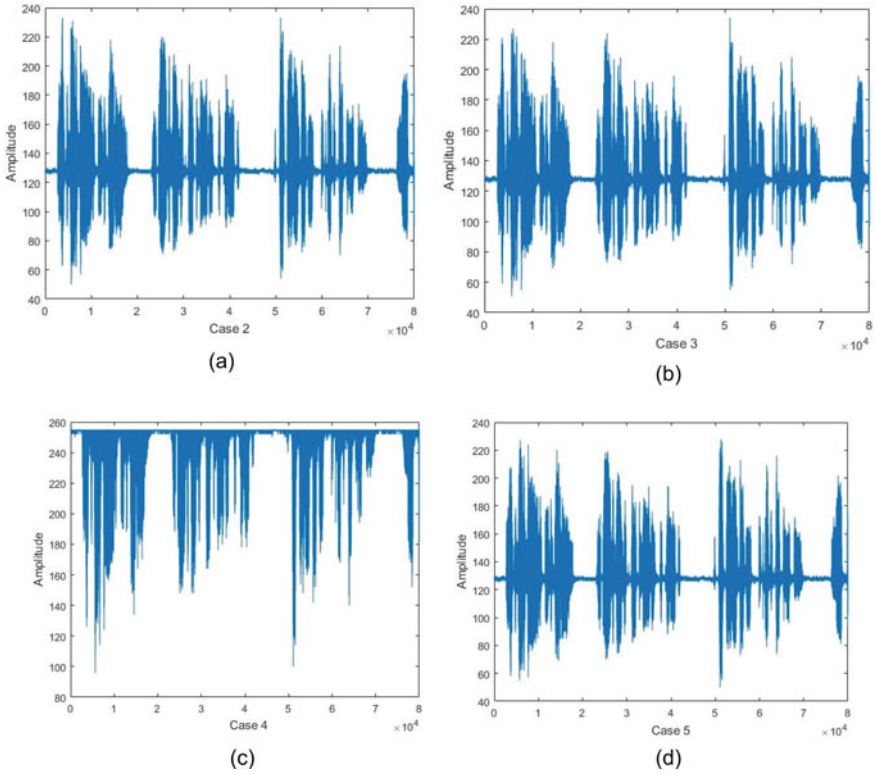


Fig. 6 **a** Case 2, **b** Case 3, **c** Case 4, **d** Case 5

Case 3, intruder attempts to enhance quality by decreasing amplitude by a fixed or random value. Figure 6b shows the low-quality share audio after intruders attempt to enhance. The correlation coefficient comes out to be 0.997 which indicates more degradation in quality than Low-Quality Share.

Case 4, intruder attempts to enhance the quality by making amplitude A_m, cA_m where c is a constant. Figure 6c shows the low-quality share audio after intruders attempt to enhance. It destroys the audio.

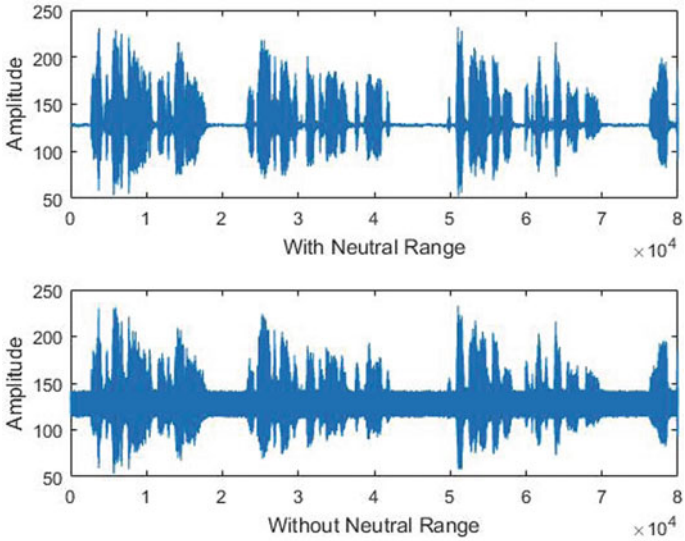


Fig. 7 Reconstructed audio without considering neutral range

Case 5, intruder attempts to enhance quality by enhancing or diminishing amplitude by a fixed or random value. Figure 6d shows the low-quality share audio after intruders attempt to enhance. The correlation coefficient comes out to be 0.9903 which indicates more degradation in quality than original Low-Quality Share.

4.4 Significance of Neutral Range

In Fig. 7 it can be seen a lot of amplitude changes, or amplitude generation around the neutral point i.e. 128. That amplitude results in a lot of unnecessary noises and can make the audio inaudible. While using a neutral range no amplitude will be generated in the neutral range, hence no generation of unnecessary noises.

4.5 Comparison with Existing ASS Techniques

There are a lot of existing schemes for Audio Secret Sharing as discussed in Sect. 2. A comparative analysis has been performed on certain parameters as shown in the Table 5.

Table 5 Comparison with existing schemes

Scheme	Correlation Coefficient = 1	Lossless	Cover Required	Quality Control
Desmedt et al. [4]	–	–	Yes	No
Ching-Nung Yang [6]	–	–	Yes	No
Prabir et al. [13]	–	–	Yes	No
Socek et al. [14]	–	–	No	No
Ehdaie et al. [7]	–	–	No	No
Vyavahare et al. [10]	Yes	Yes	No	No
Bharti et al. [11]	No	Yes	No	No
Proposed	Yes	Yes	No	Yes

5 Conclusion

In this paper, an audio quality control method based on Audio Secret Sharing has been proposed. In all the existing methods there is no technique that provides quality control on audio. This method can be used for various audio vendors for controlling the quality of their content. The CMOS, Correlation Coefficient, and PESQ(Raw MOS, MOS-LQO) come to be "Almost the same", 1 and 0.624, 1.093 respectively for original and reconstructed audio and "Slightly Worse", "0.9907" and 4.500, 4.549 respectively for original and low-quality share. Experimentally it shows the quality of the reconstructed audio and the original audio is the same and the quality of low-quality secret is degraded. This technique provides a great extent of control over quality degradation.

References

1. Shamir A (1979) How to share a secret. Commun ACM 22(11):612–613
2. Baby A, Thomas AL, Nishanthi NL, Consortium T (2016) Resources for Indian languages. In: CBBLR—community based building of language resources, Sep 2016, pp 37–43, Brno, Czech Republic: Tribun EU (Online). Available: https://www.iitm.ac.in/donlab/tts/index.php
3. Blakley GR (1979) Safeguarding cryptographic keys. In: Proceedings of American federation of information processing societies national computer conference, (AFIPS'79), California, pp 313–317
4. Desmedt Y, Hou S, Quisquater J-J (1998) Audio and optical cryptography. In: ASIACRYPT'98, LNCS, vol 1514, pp 392–404
5. Lin C-C, Laih C-S, Yang C-N (2003) New audio secret sharing schemes with time division technique. J Inf Sci Eng 19(4):605–614
6. Yang CN (2002) Improvements on audio and optical cryptography. J Inf Sci Eng 18(3):381–391

7. Ehdaie M, Eghlidos T, Aref MR (2008) A novel secret sharing scheme from audio perspective. In: Proceedings of international symposium on telecommunications (IST2008). IEEE, pp 13–18
8. Mohammad E, Taraneh E, Reza AM (2008) Some new issues on secret sharing schemes. International Conference on Telecommunications (ICT'08), June. St. Petersburg, Russia, pp 16–19
9. Wang JZ, Wu TX, Sun TY (2015) An audio secret sharing system based on fractal encoding. In: Proceedings of 49th, international Carnahan conference on security technology (ICCST), pp 211–216
10. Vyavahare S, Patil S (2016) Analysing secret sharing schemes for audio sharing. Int J Comput Appl 137(11):39–42
11. Bharti SS, Gupta M, Agarwal S (2018) A novel approach for verifiable (n, n) audio secret sharing scheme. Multimed Tools Appl 77:25629–25657
12. Naor M (1994) Visual cryptography. In: De Santis A (ed) EUROCRYPT. LNCS, vol 950, pp 1–12. Springer, Heidelberg (1995) Author F (2016) Article title. Journal 2(5):99–110
13. Naskar PK, Khan HN, Roy U, Chaudhuri A, Chaudhuri A (1996) Shared cryptography with embedded session key for secret audio., International Journal of Computer Applications (0975 – 8887) Volume 26– No.8, July 2011 ITU-T, Subjective performance assessment of telephone band and wide-band digital codecs, ITU-T Recommendation, p 830 (1996)
14. International Telecommunication Union—Radiocommunication Sector (1990) Recommendation BS. 562-3, Subjective assessment of sound quality
15. ITU (2000) Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. ITU-T Recommendation, p 862
16. Singh SK, Yadav M (2020) A (k, n) Audio secret sharing with share authentication. Int J Comput Digital Syst

Noise Removal from Audio Using CNN and Denoiser

Manmohan Dogra, Saumya Borwankar, and Jayashree Domala

Abstract As there is aggrandizement in the sector of artificial intelligence relating to the speech domain, it becomes a necessity to have efficient noise removal models with greater efficiency and less complexity. The presence of noise in audio signals poses a great complication when working on speech recognition, enhancement, improvement, and transmission. Hence, there is a necessity to develop the most efficient algorithm for noise reduction which works in real time and is successful in removing maximum noise. To be above this difficulty, this paper presents an efficient algorithm for noise detection which works on the principles of deep learning, specifically convolutional neural networks (CNNs) and the removal of similar noise from the audio using the Python module 'noise reducer.'

Keywords Noise removal · Convolutional neural network · Noise recognition · Signal to noise ratio

1 Introduction

Noise removal is a process that deals with removing noise from a signal [9]. This process can be used for both images and audios; but here, the focus is on noise removal for audios. All the signals have characteristics that make them vulnerable to noise. This noise could be white noise [11] or random noise with an even frequency distribution or frequency-dependent noise. The noise could be in the form of but not limited to hum and hiss, drillings, air conditioner, or car horn. If the audio has an unwanted and unpleasant sound, then it becomes difficult to use it further for transmission, recognition, or improvement. Considering the various edge devices like phones, laptops, conferencing systems, etc., traditional noise suppression methods

M. Dogra (✉) · J. Domala
St. Francis Institute of Technology, Mumbai, Maharashtra 400103, India

S. Borwankar
Nirma University, Ahmedabad, Gujarat 382481, India
e-mail: 17bec095@nirmauni.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_4

have been implemented. Since it is an edge device, such an approach seems intuitive. It captures the user's voice in the first place, and then device filters are used to noise out the captured voice. Later, the output is saved [6]. Another common and consistent solution for controlling or suppressing background noise is sound masking. It adds the white noise commonly referred to as the pink noise into an environment to mask the unwanted sound. But this method is not an active noise control. That is, it only reduces or eliminates the perception of noise [10].

As a deduction to this issue at hand, 'Noise removal from audio using CNN and denoiser' is introduced as an affirmative solution to solving the problems that arise due to the presence of noises in the audio. A broad goal that we wish to accomplish is noise removal to the bare minimum with the accessibility of doing it in real time. The core of the project is based on the deep learning principles of convolutional neural networks (CNNs). CNN is a special type of artificial neural network (ANN) which has the ability to pick the patterns and makes sense out of it. These networks take advantage of local spatial coherence which helps them have shared parameters leading to fewer weights, thereby forming convolutions [8]. This aids in extracting relevant information and reduces the computational cost. The system will apply the deep learning algorithm to detect the type of noise in the audio and eventually, a 'noise reducer' will be used to remove that particular noise to finally produce a noise-free audio output.

2 Relevant Work

Various researchers have worked on noise reduction techniques utilizing different methodologies. The authors Chavan O., et al., have worked on noise reduction with the help of audacity noise removal. This is an open-source audio program. The decrease of clamor was achieved by using the Fourier analysis to improve the audio frequency spectrum. The spectrum of pristine tones that make up the background noise found in the susurration segment [7]. In another paper by Al-Allaf, et al., the authors have made four ANN models. These models are namely function fitting (FitNet), recurrent neural networks (RNNs), cascaded-ForwardNet, and nonlinear autoregressive (NARX). They have prepared them independently to turn into a channel to dispose of the noise from any speech signal. From tests led, the best outcomes were gotten from NARAX and FitNet models. But the best algorithm in terms of training, in this case, is TrainLM [4]. While in the paper written by Zoican S., the author presents a framework dependent on the wavelet approach with a level-dependent threshold that gives a valuable technique to eliminate noise in the signals. The test results show that this technique eliminates noise altogether and the framework works progressively [19]. The authors Prasadh S.K., et al., have utilized the strategy where the noise signal is passed through a channel or transformation. Signal dependent rank order mean algorithm (SD-ROM), which eliminates noise from audio signals, is used. It also retains the characteristics of the signal. To wipe out white Gaussian noise, a discrete wavelet transform method is used. After every one

of the methods is applied to the examples, SNR and elapsed time are calculated. The entirety of the above methods shows an increased signal to noise ratio (SNR) after processing [17]. A paper by Godsill S., et al., the authors have introduced a signal separation-based way to deal with this issue. Noise transients and audio signals are displayed as autoregressive (AR) measures which are superimposed additively to give the observed waveform. A more practical scheme is then evolved which uses a Kalman filter for the implementation of the separation. The excitation variance of the noise transient model is tightened exponentially to zero away from the irregularity to evade low-frequency distortions to the sound signal. The strategy is totally mechanized and, moreover, pragmatic to actualize when contrasted with the current plans for the removal of such deformities. Results indicate a high level of performance [14]. Authors Germain F.G., et al., present a deep learning approach to deal with the process of diminishing the noise in the speech signals by processing the crude waveform directly. The framework plans to deliver a processed signal that contains only the speech content for a given input audio. This input audio would contain speech tainted by an additive noisy background signal. A fully convolutional context aggregation network is trained using a deep feature loss. This deep loss feature loss is based on the comparison. The juxtaposition is done between the internal feature activation in a different network, trained for acoustic condition identification and domestic audio tagging [13]. The writers Murphy J., et al., give a technique for the disturbance of commotion which includes non-Gaussian impulse from a signal. Gabor relapse version helps in the removal of impulse noise. The parameters are calculated by Gibbs MCC sampler of problem formulated by the joint Bayesian framework. Signal to noise ratios improved, and the audio quality of the result improved by modeling impulses with the discrete switching process [16][18]. While in another paper, the authors Ali M., et al., used a hierarchical threshold algorithm which presented noise removal based on double-intensity dual-tree wavelet transform (DDDTWT). Additive white Gaussian Noise mixed with the audio signal is selected for the implementation. The results produced by the global threshold method signal to noise ratio, DDDTWT, and root mean square error (RMSE) are observed and juxtaposed[5]. Considering the previous research already done in this area, the proposed system focuses more on deep learning-based technique, specifically using CNN. The main goal is to implement a more convenient and sophisticated system that is reliable.

3 Implementation

The implementation process starts by taking in the input noisy audio which goes through the deep learning model to detect the type of noise present in the input audio. Then, the input noisy audio and the detected noise type are passed through a 'noise reducer' module of Python which damps the noise from the original noisy audio to give a noise suppressed audio with less audio turbulence. The noisy audio dataset used for training the neural network is by 'UrbanSound8k' 6GB in size containing 8732 sound excerpts which are labeled of from ten most common noise classes:

children, air conditioner, dog bark, drilling, car horn, engine running, jackhammer, gunshot, street music, and continuous siren. The pseudo-code for implementation is as follows:

Algorithm 1 Model creation

```
Require: N is the total number of noisy training audios
procedure MATRIX(noisy audio)                                   ▷
    A1 ← MFCC(audiofile)                                        ▷ Generate MFCC array A1
    A2 ← Melspectrogram(audiofile)                              ▷ Generate mel spectrogram array A2
    A3 ← chromaSTFT(audiofile)                                  ▷ Generate chroma STFT A3
    A4 ← chromaCQT(audiofile)                                   ▷ Generate chroma CQT array A4
    A5 ← chromaCENS(audiofile)                                  ▷ Generate chroma CENS array A5
    S1 ← A1, A2, A3, A4, A5                                     ▷ Stacked to form array S1
end procedure
for i in range N do
    append Si to Sn
    Repeat the procedure from step 1
end for
for i in range N do
    Si converted to one-hot encoded vector Pi
    Vi padded to form Pi
    Pi reshaped to form Ri
end for
procedure MAKE MODEL                                            ▷ Creation of model
    while layers do
        model ← model.layers                                    ▷ Defining model architecture
    end while
    model ← model.compile                                       ▷ Compiling the model
    return model
end procedure
procedure TRAINING        ▷ Model is trained on the vectors Ri and clustered in classes Ci
end procedure
save model.h5                          ▷ This model file can be used to make predictions
procedure LOADING MODEL                ▷ Model M is loaded with noisy input audio X1
    X1 undergoes from step 1 to step 17 giving finally padded P1
    P1 is passed through Model M, which predicts the noise class
  C1 and C2
    Noise N1 of type C1 and N2 of type C2 is fetched from noise
  samples.
end procedure
procedure NOISE REDUCER                ▷ Using the library noise reducer to remove noise
    noise reducer(X1,N1)
    noise reducer(X1,N2)
end procedure
```

This implementation is broken down into three modules. These modules are mainly pre-processing, training, and noise removal.

3.1 Module 1. Pre-processing

This module deals with extracting the relevant information from the noisy audio which is later used for training purposes. The audio has five different features on which we mainly focus. These features are:

1. MFCC: Mel frequency cepstral coefficients—It has 39 capabilities, and the function be counted is small enough to pressure us to examine the records of the audio. 12 parameters are associated with the amplitude of frequencies. It affords us with sufficient frequency channels to research the audio[15].
2. Mel spectrogram: Compute a mel-scaled spectrogram[3]—It samples the input with window size and computes the fast Fourier transform for each window from the time domain to the window domain. Takes the entire spectrum and evenly separates it into spaced frequencies and generates spectrograms [12].
3. Chroma STFT: Chroma short-term Fourier transfer—Compute a chromagram from a waveform or power spectrogram using short-term Fourier transforms [2].
4. Chroma CQT: Constant-Q transform—Like mel-scale, the regular-Q transform uses a logarithmically spaced frequency axis, where the window period is distinctive for every frequency, i.e., low frequency for long windows and high frequency for short window [1].
5. Chroma CENS: Chroma energy normalized statistics: The CENS features take statistics facts over huge windows smooths local deviations in tempo, articulation, and musical embellishes inclusive of trills and arpeggiated chords. CENS are best used for tasks such as audio matching and similarity [1].

The more number of features, the better is the collection of data for training the model. Therefore, these five features are taken into account.

The below steps go around in a loop as long as the audio is inputted.

1. Step 1: Input audio. We get the noisy audio input which is in the 'wav' format.
2. Step 2: Generate feature array. The five features mentioned above are extracted from the noisy audio and are stored in an array. These arrays generated for each noisy audio are further saved in a csv file.

3.2 Module 2. Training

This module deals with the core deep learning algorithm which detects the type of noise present in the input audio.

1. Step 1: Retrieve the amalgamated array from the csv file—The csv file, which has each input audio's extracted features information, is used. The csv file is retrieved in the form of a dataframe.
2. Step 2: Convert to categorical features—The algorithm works efficiently if the input values are converted to categorical values. This is done by using one-hot encoding. Each array is converted into a one-hot encoded vector.

3. Step 3: Padding—The next step is to apply padding to these one-hot encoded vectors. This is done to make the length of the vectors equal.
4. Step 4: Reshaping—Since the deep learning algorithm used is convolutional neural networks (CNNs), it needs to be reshaped into the 2D form. This final matrix is then stored in the data frame. The matrix in 2D form is retrieved and then converted in 3D form before training.
5. Step 5: Building the model—The model used has 200 input variables. Moreover, it is formed by two 'Convolutional' layers: first with 128 neurons and the second layer with 256 neurons, both with a kernel size of 5, the number of strides as 1 with 'relu' as the activation function. Then, following are the two 'MaxPooling' layers with the same padding and then a dropout layer of 0.3 removing 30% of neurons, followed by a flatten layer. The flatten layer has 5120 variables. The flatten layer is followed by a hidden(dense) layer with 512 neurons and 0.3 dropout layer and then a subsequent dense layer with 1024 neurons and 0.3 dropout layer. In the end, an output layer of ten categorical output is set. A total set of trainable parameters (3,980,298) was compiled using 'adam' optimizer and 'categorical_crossentropy' as the loss function. The training took place with a batch size of 50 and 40 epochs. The layers of the CNN architecture can be seen in Fig. 2.
6. Step 6: Training the model—The last step is the training of the model. The model is trained on the 'train' and 'validation' sets of the data frame and passed through the convoluted neural network. The trained model is saved for future purposes of testing. The flowchart representing pre-processing and training steps is shown in Fig. 1.

The problem with the reduction of noise is the recognition of noise; if the noise signal is directly available to us, it becomes very easy to remove the noise, but in real-case scenarios, the noisy audio is not present separately; instead, it is mixed with

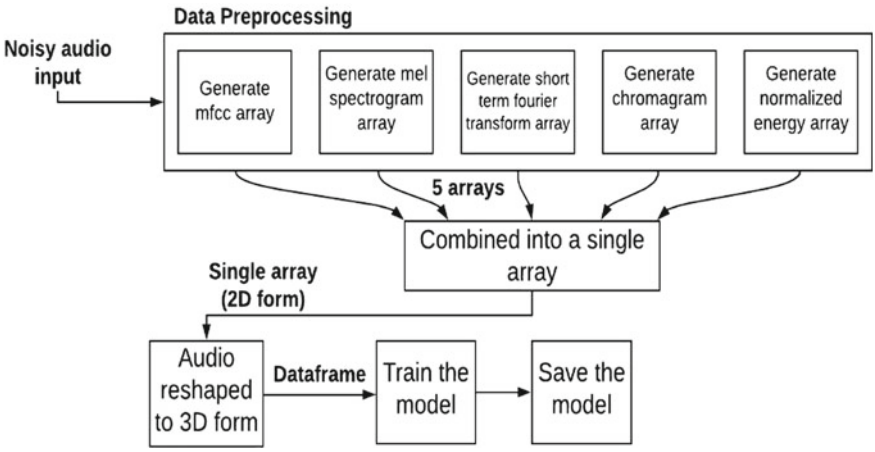


Fig. 1 Flow diagram of the pre-processing and training module

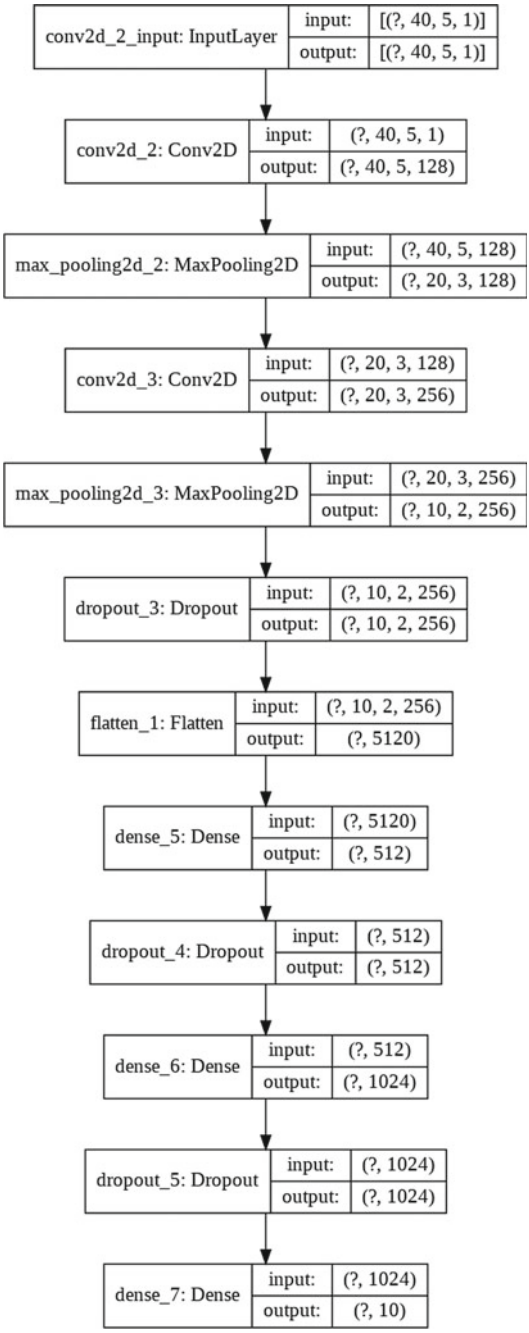


Fig. 2 Representation of the layers of the CNN architecture

the target input which creates a problem to localize the noise. Our method helps to tackle this problem as it provides us a way to recognize the noise clip before hand and later removes the noise with the help of two separate clips first being the noise + original audio clip and the second being the noisy audio clip. So, this helps us to convert noisy input to a denoised signal with the help of steps described in module 3 Step 4, noise reduce library of Python helps us achieve the last task.

3.3 Module 3. Noise Removal

This module deals with the final de-noising part.

1. Step 1: Input audio and load model—The previously trained model and the audio input are loaded.
2. Step 2: Pre-processing—The input audio goes through the pre-processing steps as mentioned in module 1. This pre-processing step extracts the features of audio and presents in the 3D vector form.
3. Step 3: Noise type detection This 3D vector is then sent through the trained model. This model outputs the type of noise detected. The detection takes place by comparing a predefined noise dataset with the input noisy audio. This comparison helps to retrieve the types of noise present in the audio.
4. Step 4: Denoiser After the successful detection of the noise, the detected type of noise and original input audio are sent through the 'noise-reducer' Python module. This module removes the noise from the audio and generates a less turbulent audio having reduced noise. The Python module is able to do certain transformations. Firstly, FFT is calculated on the noise clip, after which the statistics are computed over the FFT of the noise. Next, a threshold is set according to the statistics of the noise. Now, the FFT of the signal is calculated, after which a mask is calculated by the comparison of the threshold to the signal FFT. Next, the mask is smoothed with a filter over time and frequency. Lastly, the mask is put on the FFT of the signal and is inverted (Fig. 3).

4 Results

Evaluation metrics play an important role to ascertain the reliability of the model. For the CNN model, the accuracy and loss can be determined which reflects the efficiency of the model. The CNN model gives the training accuracy as **95.82%**, training loss as **0.13**, validation accuracy as **73.59%**, and validation loss as **1.02** as shown by the graphs in Fig. 4.

To check the model output, two samples of noisy data are taken and passed through the CNN model after which the noise reduction takes place in the noise reduction package. Figure 5 represents a noisy input with a street background noise.

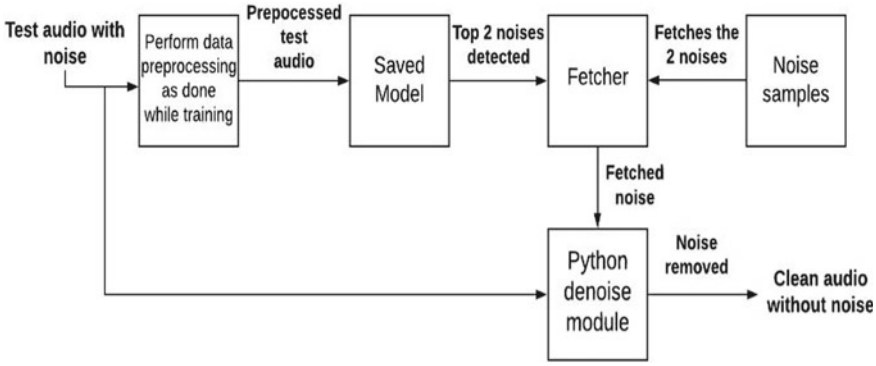


Fig. 3 Flow diagram of the noise removal module

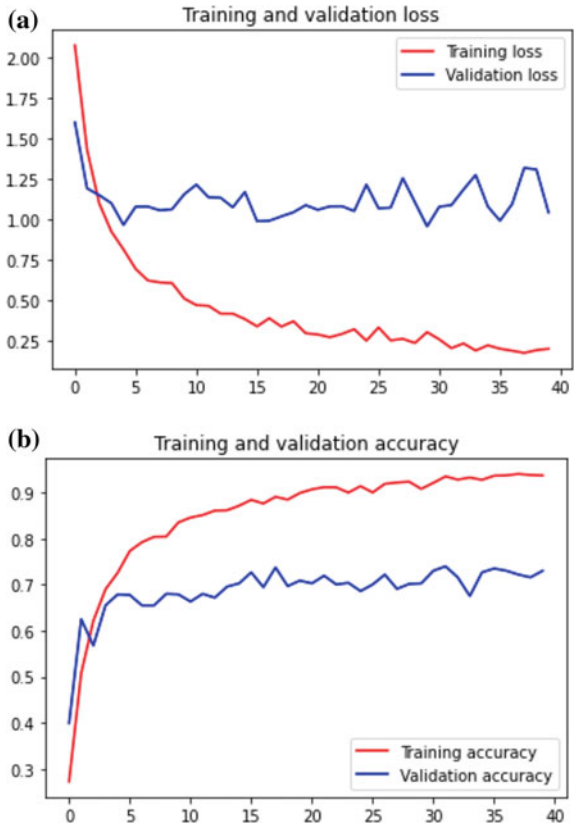


Fig. 4 Training and Validation **a** Loss and **b** Accuracy on dataset

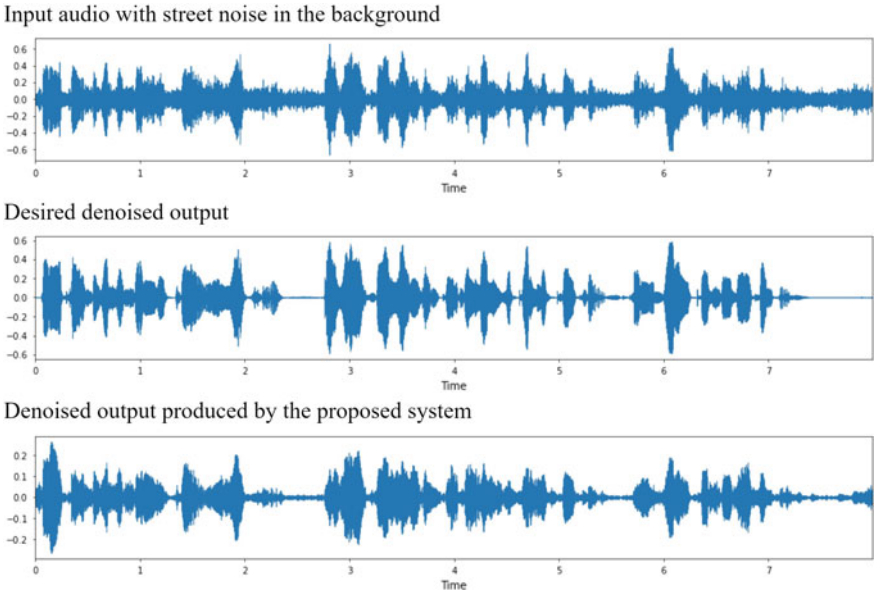


Fig. 5 Noise reduction output #1

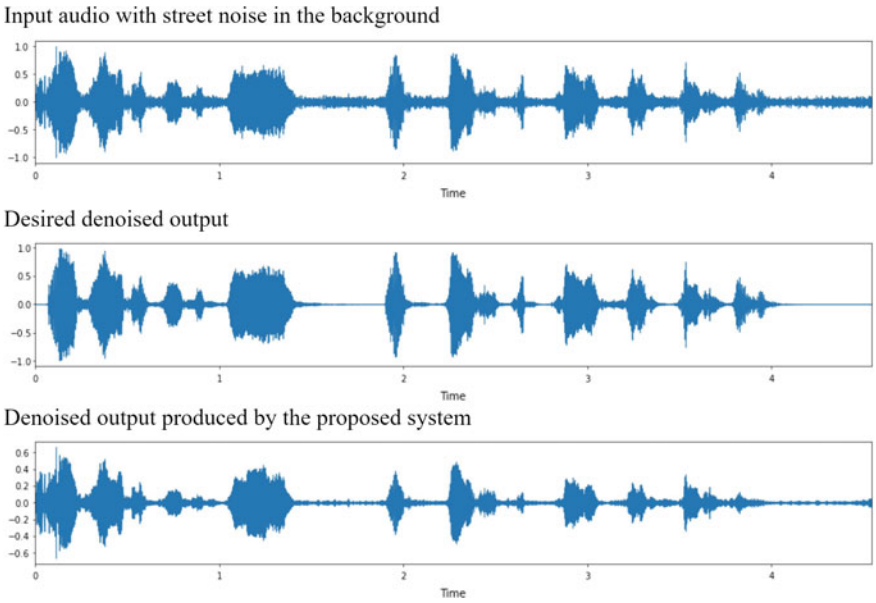


Fig. 6 Noise reduction output #2

The desired output, along with the output generated by the system, are also shown. Similarly, Fig. 6 represents a noisy input with children's background noise along with the desired and generated output.

5 Conclusion

Through this paper, a system for noise removal from the audio was presented. In particular, a deep learning-based model of CNN was used to detect the noises and then perform the noise removal from audio. The primary research aim was to build an extrinsic and efficient model that can be further incorporated for any audio-related use case. The experimental results show the excellent performance of the model, which achieved an accuracy of 97.1%.

For the future scope of this project, an increase in the number of features to get more accurate results and applying HQ filters to mask the noise in audio and prevent audio-loss can be done.

References

1. Constant-q transform and chroma. https://musicinformationretrieval.com/chroma.html
2. librosa.feature.chroma_stft. http://man.hubwiz.com/docset/LibROSA.docset/Contents/Resources/Documents/generated/librosa.feature.chroma_stft.html
3. librosa.feature.melspectrogram. http://man.hubwiz.com/docset/LibROSA.docset/Contents/Resources/Documents/generated/librosa.feature.melspectrogram.html
4. Al-Allaf O (2015) Removing noise from speech signals using different approaches of artificial neural networks. Int J Information Technol Comput Sci 7:8–18. https://doi.org/10.5815/ijitcs.2015.07.02
5. Ali MA, Shemi PM (2015) An improved method of audio denoising based on wavelet transform. In: 2015 international conference on Power, Instrumentation, Control and Computing (PICC), pp 1–6
6. Baghdasaryan D (2020) Real-time noise suppression using deep learning. https://developer.nvidia.com/blog/nvidia-real-time-noise-suppression-deep-learning/
7. Chavan O, Kharade N, Chaudhari A, Bhalke N, Nimbalkar P (2019) Machine learning and noise reduction techniques for music genre classification. Mach Learn 6(12)
8. Contributors W (2020) Convolution; wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Convolution&oldid=975151751. Online. Accessed 6 Sept 2020
9. Contributors W (2020) Noise reduction, wikipedia the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Noise_reduction&oldid=967676491. Online. Accessed 6 Sept 2020
10. Contributors W (2020) Sound masking; wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Sound_masking&oldid=965358994. Online. Accessed 6 Sept 2020
11. Contributors W (2020) White noise; wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=White_noise&oldid=975542178 Online. Accessed 6 Sept 2020
12. Gartzman D (2020) Getting to know the mel spectrogram. https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0
13. Germain FG, Chen Q, Koltun V (2018) Speech denoising with deep feature losses

14. Godsill SJ, Tan CH (1997) Removal of low frequency transient noise from old recordings using model-based signal separation techniques. In: Proceedings of 1997 workshop on applications of signal processing to audio and acoustics, p 4
15. Hui J (2019) Speech recognition-feature extraction mfcc & plp. https://medium.com/@jonathan_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9
16. Murphy J, Godsill S (2011) Joint bayesian removal of impulse and background noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 261–264
17. Prasadh SK, Natrajan SS, Kalaivani S (2017) Efficiency analysis of noise reduction algorithms: analysis of the best algorithm of noise reduction from a set of algorithms. In: 2017 International Conference on Inventive Computing and Informatics (ICICI), pp 1137–1140
18. Wolfe P, Godsill S, Ng WJ (2004) Bayesian variable selection and regularization for time-frequency surface estimation. J Royal Stat Soc Ser B 66:575–589. https://doi.org/10.1111/j.1467-9868.2004.02052.x
19. Zoican S (2010) Audio signals noise removal real time system, pp 25–28. https://doi.org/10.1109/ICCOMM.2010.5509098

Sine-Wave Speech as Preprocessing for Downstream Tasks

Sri Harsha Dumpala, Jason d'Eon, and Sageev Oore

Abstract Analyzing raw speech signals is computationally expensive. Hence, a set of simplified speech representations is proposed for downstream tasks—such as emotion recognition—that do not depend on vocal timbre or textual content. These representations are obtained by considering sinusoids replicating the frequency and amplitude of the resonance peaks of speech. We show that for a representation based on combining several such peaks, untrained individuals are often able to recognize emotive states in the sinusoidal representation corresponding to the same emotive states of the original speech signal.

Keywords Sine-wave speech · Speech · Emotion · Formant frequencies

1 Introduction

Speech is a complex signal produced by excitation of a time-varying vocal tract system [6]. Analysis of speech signals has long been of great interest for many applications, such as text-to-speech synthesis [10], modifying the time-scale and pitch of a speech recording [5, 12], recognizing the speaker [4], and classification of emotion and even detection of clinical depression by voice [14, 20].

Raw, unprocessed speech signals are computationally expensive to analyze, so many attempts have been made to obtain a simplified representation of speech signals [8, 13]. By Fourier analysis, any interval of the time-frequency representation of a speech signal can be interpreted as the sum of sine-wave frequencies over that

S. H. Dumpala and J. d'Eon: Both authors contributed equally

S. H. Dumpala (✉) · J. d'Eon · S. Oore
Dalhousie University, 6299 South Street, Halifax, NS, Canada
e-mail: js348697@dal.ca

Vector Institute, 661 University Avenue, Toronto, ON, Canada
e-mail: sriharsha.d@dal.ca

S. Oore
e-mail: sageev@dal.ca

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_5

interval. Additionally, it is known that much of the linguistic information in speech can be reconstructed by locating local maxima of the spectral envelope (formants) which are caused by resonances in the vocal tract [6]. Sine-wave speech simplifies the speech signal by taking only the sum of two or three time-varying sinusoids that track the frequency and amplitude estimates of the first few formant frequencies [17].

In the initial study introducing sine-wave representation of speech, the linguistic content in the sine-wave representation as perceived by human listeners was analyzed [17]. It was evident from these perceptual tests that naive listeners (with no prior knowledge about the nature of the sounds) were unable to recognize the sine-wave representation as even speech. But listeners with prior knowledge about the identity of the original speech signal were able to relate the sine-wave to the linguistic content with high confidence. Interestingly, the majority of the naive listeners recognized the sine-wave representations as music or whistling by humans or rhythmic computer beeps. Perceptual identification of the individual speakers from sine-wave speech was also analyzed [16, 18, 19]. It is evident from these studies that the listeners were able to identify the individual speakers, even with the acoustic correlates of vocal tract scale variation removed [16]. Sine-wave speech has previously been studied to determine its content of phonetic information for English and tonal languages such as Mandarin [3, 7], and has even been tested in transcription tasks [2]. In this paper, we explore the perceptual recognition of emotions from sine-wave speech.

Perception of emotion in speech by humans has long been a topic of research interest [9, 15], but it was not clearly evident from these studies how the linguistic content of the speech signals influences the perception of emotion by humans. This study might provide useful insights into this research topic as sine-wave speech is mostly devoid of linguistic content and mainly represents prosodic patterns of the speech signal.

In this study, we analyzed how the sine-wave speech representation affects the perception of emotion in speech. Two questions that we investigate in this study are:

- Can the sine-wave speech synthesized from a natural speech utterance retain emotional characteristics of that utterance?
- How does the combination of different resonant frequencies (F_1, F_2 and F_3) affect the perception of emotion by listeners?

2 Sine-Wave Representation of Speech

The sine-wave representation of a speech signal is obtained by combining multiple sinusoids, whose instantaneous frequencies and amplitudes are conditioned on the estimated instantaneous frequencies and amplitude patterns of the resonance peaks of a natural speech signal [17]. In this study, we will consider the first three resonant frequencies (formant frequencies). Our approach to extract sine-wave representation of speech is detailed in the following steps:

1. We estimate the magnitude spectrogram (time-frequency representation) of the speech signal (sampled at 16 kHz) by computing the short-time Fourier transform (1024 point Fourier transform) of the speech signal using a frame size of 50 ms and a frame shift of 12.5 ms.
2. We estimate the first three instantaneous formant frequencies (i.e., F_1, F_2 and F_3) of the speech signal within a window of size 50 ms (frame length = 50 ms) for every 12.5 ms (frame-shift = 12.5 ms).
3. We consider the instantaneous amplitude of the magnitude spectrogram at the formant frequency as the amplitude of the formant frequency.
4. We obtain an energy-based voiced/unvoiced decision on each frame of the speech signal. We use a threshold of 0.05 for the voiced/unvoiced decision. We retain only the voiced frames and discard the unvoiced frames as the formant estimation is not reliable in the unvoiced regions.
5. We generate time-varying sine-waves whose instantaneous frequency is equal to the frame-level formant frequencies, and the amplitude of the sine wave is the amplitude of the formant frequencies (computed in step 3). Sine-F_1 refers to the sine wave generated based on first formant frequencies (F_1). Similarly, Sine-F_2 and Sine-F_3 refers to the sine waves generated based on F_2 and F_3, respectively.
6. Combined sine waves are obtained by simply adding the sine waves corresponding to a particular formant frequency. For instance, Sine-$F_1F_2F_3$ is obtained by simply taking the sum $\sum_{i=1}^{3}$ Sine-F_i.

3 Experiments and Results

3.1 Dataset Used

In this work, we used the EmoV-DB: The Database of Emotional Voices dataset [1]. The EmoV-DB dataset consists of sentences uttered by 3 actors (one male and two female) in English. Each actor uttered a subset of the sentences from the CMU-Arctic [11] database, which are phonetically balanced. The actors uttered these sentences with five different emotion variations (neutral, amusement, anger, disgust, and sleepiness) in parallel, i.e., the same sentence spoken in different emotive states. In this study, we considered 3 different utterances spoken by one female actor in three different emotive states (amusement, anger, and sleepiness).

3.2 Experimental Studies

We performed perceptual evaluation on the sine-wave representations obtained for the spoken utterances in different emotive states. Three different sine wave combinations are considered: (a) Sine-F_1, (b) Sine-F_3 and (c) Sine-$F_1F_2F_3$.

A total of 14 subjects participated in the perceptual evaluation tests. The listeners were asked to annotate the samples based on the emotion perceived by the listener. In these experiments, the listeners did not have any prior knowledge about the linguistic content of the speech signal other than the fact that these sinusoids were derived from speech signals. In order to make the listeners familiar with the evaluation criteria, the listeners were provided with several sample speech utterances and their corresponding sine-wave chirps extracted from the samples. Each listener annotated a total of 27 samples (3 utterances $\times$ 3 emotions $\times$3 sine representations). The samples are randomized for every utterance and for every subject to avoid any bias to the order of the occurrence of the samples.

3.3 Results

Results of the perceptual evaluation are shown in Figs. 1 and 2. We show the predictions over the entire experiment, as well as confusion matrices for each of Sine-$F_1F_2F_3$, Sine-F_1, and Sine-F_3.

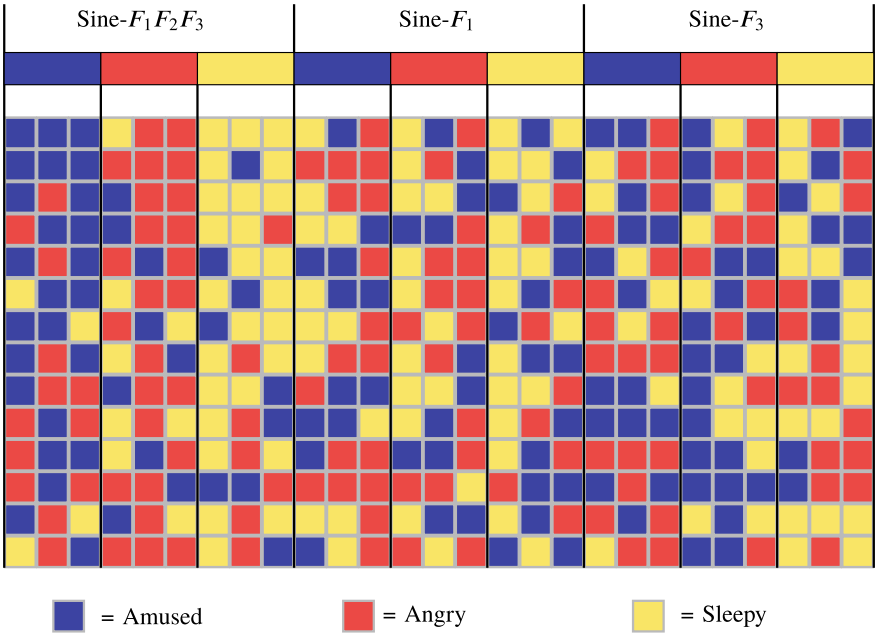


Fig. 1 Predictions made by participants in the perceptual evaluation. Each row of the grid represents one participant, and each column represents one question in the experiment. The colors in the grid show what the predicted classification was, and the bars at the top correspond with the true class

Actual \ Predicted	Amused	Angry	Sleepy
Amused	25	13	4
Angry	9	25	8
Sleepy	9	7	26

(a) Sine-$F_1F_2F_3$

Actual \ Predicted	Amused	Angry	Sleepy
Amused	12	18	12
Angry	12	16	14
Sleepy	16	9	17

(b) Sine-F_1

Actual \ Predicted	Amused	Angry	Sleepy
Amused	15	20	7
Angry	20	11	11
Sleepy	10	14	18

(c) Sine-F_3

Fig. 2 Confusion matrices obtained from perceptual evaluation of emotion using sine-wave speech consisting of **a** first three formants, **b** first formant only, and **c** third formant only, respectively

4 Discussion

By the results in Fig. 2, we can see that using a combination of the first three formants allows each emotive state to be correctly predicted nearly two thirds of the time. Not surprisingly, using F_1 or F_3 alone leads to higher confusion between certain states. Moreover, the small sample size of this exploratory study prevents us from drawing any particular conclusions. However, taken together, these results do show that while single formants may not be sufficient for a human to accurately do emotion classification, the sum of all three formants *can* be sufficient to do so.

We were also interested in seeing how a particular samples in the survey had highly consistent answers, even if those answers were wrong. Figure 1 provides a more detailed look at classification performance. Again, it shows that combining formants made the task easier, but we can also see that the third column in Sine-F_1 had a large majority of participants answer "Angry" when the truth was "Amused". On the other hand, the fourth column in Sine-F_3 has mostly "Amused" answers when the truth was "Angry". This suggests that when only one formant is used, these specific emotions can sound like one another.

5 Conclusions and Future Work

We have articulated a clear methodology for converting speech to a concise sine-wave representation based on formant frequencies. While doing this conversion removes linguistic content of speech and timbral characteristics of the voice, we have indicated that this approach can maintain enough prosodic elements that it enables humans to still recognize certain emotive states in speech, that is, effectively, no longer recognizable as speech. This suggests that sine-wave representation of speech may prove to be a useful preprocessing step for potential downstream tasks and applications, including emotion recognition.

Acknowledgements The authors would like to thank the reviewers for providing their valuable comments which helped in improving the quality of the paper. The authors would also thank the Government of Canada for supporting this research through CIFAR.

References

1. Adigwe A, Tits N, Haddad KE, Ostadabbas S, Dutoit T (2018) The emotional voices database: Towards controlling the emotion dimension in voice generation systems. arXiv preprint arXiv:180609514
2. Barker J, Cooke M (1999) Is the sine-wave speech cocktail party worth attending? Speech Commun 27(3–4):159–174
3. Benson RR, Richardson M, Whalen DH, Lai S (2006) Phonetic processing areas revealed by sinewave speech and acoustically similar non-speech. Neuroimage 31(1):342–353
4. Dumpala SH, Kopparapu SK (2017) Improved speaker recognition system for stressed speech using deep neural networks. In: 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1257–1264
5. Dumpala SH, Sheikh I, Chakraborty R, Kopparapu SK (2018) Cycle-consistent gan front-end to improve asr robustness to perturbed speech. In: NeurIPS workshop on interpretability and robustness in audio, speech, and language
6. Epps J, Smith JR, Wolfe J (1997) A novel instrument to measure acoustic resonances of the vocal tract during phonation. Meas Sci Technol 8(10):1112–1121. https://doi.org/10.1088/0957-0233/8/10/012
7. Feng YM, Xu L, Zhou N, Yang G, Yin SK (2012) Sine-wave speech recognition in a tonal language. J Acoustical Soc Am 131(2):EL133–EL138
8. Hermansky H (1990) Perceptual linear predictive (plp) analysis of speech. J Acoustical Soc Am 87(4):1738–1752
9. Izard CE (2013) Human emotions. Springer Science & Business Media
10. Klatt DH (1987) Review of text-to-speech conversion for English. J Acoustical Soc Am 82(3):737–793
11. Kominek J, Black AW (2004) The cmu arctic speech databases. In: Fifth ISCA workshop on speech synthesis
12. Macon MW, Clements MA (1997) Sinusoidal modeling and modification of unvoiced speech. IEEE Trans Speech Audio Process 5(6):557–560
13. Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63(4):561–580
14. Ooi KEB, Lech M, Allen NB (2012) Multichannel weighted speech classification system for prediction of major depression in adolescents. IEEE Trans Biomed Eng 60(2):497–506
15. Plutchik R (1984) Emotions: a general psychoevolutionary theory. Approaches to Emotion 1984:197–219
16. Remez RE (2008) Sine-wave speech. Scholarpedia 3(7):2394. https://doi.org/10.4249/scholarpedia.2394, revision #73091
17. Remez RE, Rubin PE, Pisoni DB, Carrell TD (1981) Speech perception without traditional speech cues. Science 212(4497):947–949
18. Remez RE, Fellowes JM, Rubin PE (1997) Talker identification based on phonetic information. J Exp Psychol Human Percept Performance 23(3):651
19. Sheffert SM, Pisoni DB, Fellowes JM, Remez RE (2002) Learning to recognize talkers from natural, sinewave, and reversed speech samples. J Exp Psychol Human Percept Performance 28(6):1447
20. Ververidis D, Kotropoulos C, Pitas I (2004) Automatic emotional speech classification. In: 2004 IEEE international conference on acoustics, speech, and signal processing. IEEE, vol 1, pp I–593

Style of Vocal Singers in Indian Classical Music: Timbre Approach

Anirban Patranabis, Kaushik Banerjee, Ranjan Sengupta, and Dipak Ghosh

Abstract Can the sound of a vocalist be qualified by timbre? Can we identify the style of a vocalist from timbre? In this work, our objective is to identify timbre characteristics of vocal sound signals of eminent vocalists in Hindustani music, and we tried to identify the style of the artist from those timbre structures. The descriptive timbre parameters that help with identifying the style of vocalists are brightness, odd and even harmonics, irregularity among partials and spectral centroids. Beside these, shimmer, jitter and frequency per partial number were also descriptive. The study concludes that timbre is an important feature to distinguish or identify the style of a vocalist.

Keywords Hindustani vocal music · Timbre parameters · Spectral parameters · Style of vocalist

1 Introduction

Vocal performers developed unique styles by developing elements of vocal style (*gayaki*) by practice and from their experience with the listener's preferences. A musical sound must have three main dimensions: production, acoustic signal properties and perception [1]. This paper discusses the timbre aspect (production aspects) of vocal sound along with their acoustic analyses. The inter-relationships of the acoustic attributes and subjective response to music span vast areas are hearing perception, musicology, aesthetics and psychology [2]. Timbre analysis bridges these inter-relationships.

A person versed in Indian classical music can identify the style of a singer, contrary to a lesser experienced listener. In this study, we have computationally conducted timbre analyses of the sound signals of eminent vocalists in Hindustani music. We aimed at the identification of the style of the vocalists based on timbre characteristics of these musical signals. Among timbre parameters, brightness [3], tristimulus [4], odd and even harmonics [5], irregularity among partials [2, 6] and spectral centroids

A. Patranabis (✉) · K. Banerjee · R. Sengupta · D. Ghosh
Sir C V Raman Centre for Physics and Music, Jadavpur University, Kolkata, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_6

[7] were studied. Among spectral features, shimmer and jitter [8] were also studied. This approach leads to an understanding the styles of the artists based on the timbre changes.

We focused only on the spectral and temporal characteristics of the musical sounds. The temporal sequence of change is considered as the timbral structure. We found that some sequences of these timbre structures became a crucial information carrier of sound classification and stylistic identification of vocalists in Hindustani music. For this study, we have chosen twenty eminent singers. Each of them sung four ragas, namely Bhairav, Todi, Darbari Kannada and Mian-Ki-Malhar. So, eighty songs were taken for analysis. Only the vocal part of a rendition without any influence of accompanying instruments like the "tabla" (known as aalap) of ragas sung by eminent singers in Indian classical music was considered. Digitization of the signal was done at 44,100 samples/second (16 bits/sample). Partial to partial characteristics were studied from long-term average spectra (LTAS) of each signals, and these constitute the database for the study.

2 Results and Discussion

Additive sinusoidal components obtained from the harmonic sounds are called partials. Those partials have time-varying amplitude and frequency. The frequencies of the partials are most likely multiples of the fundamental frequency. The fundamental frequency is measured in the frequency domain using a fast Fourier transform (FFT). Partial to partial characteristics that include mainly the frequency and amplitude of each partial were studied from LTAS of each signal, and all the timbre parameters were computed. Figures 1, 2, 3, 4 and 5 show only the similar timbre parameters for individual artists.

Uniqueness in timbral qualities of the singers ABD, NIR and SUC is observed in T1, T3 and odd–even partials. Fundamental and high frequency partials are two unique timbral characteristics of these artists. Spectral energy peaked in the high frequency partials. Dissimilarity is observed in midrange frequencies. Other timbre parameters like brightness, centroid and irregularity show little difference in four renderings. Uniqueness in timbral qualities for the singer AJO, WAS and ANI is observed in T3 and odd–even partials. High frequency partials are the unique timbral characteristics of these artists as the energy is peaked up at high frequency partials. Odd and even partials of voices are identical for all these five singers. Other timbre parameters show much difference in four renderings. Therefore, timbre dissimilarities are observed for these three artists.

Timbre similarity is highly visible in the renderings of the artist ARU, ARS, AMI, ATK, FAL, SAN and MAS. In these signals, parameters like T1, T2, T3, spectral irregularity and odd–even relations are similar. Energy levels of all frequency partials are higher. Odd and even partials are almost overlapping each other. Uniqueness lies in the fundamental, mid-frequency and high frequency partials. The variations of frequency components are very low in all the renderings, and hence almost no

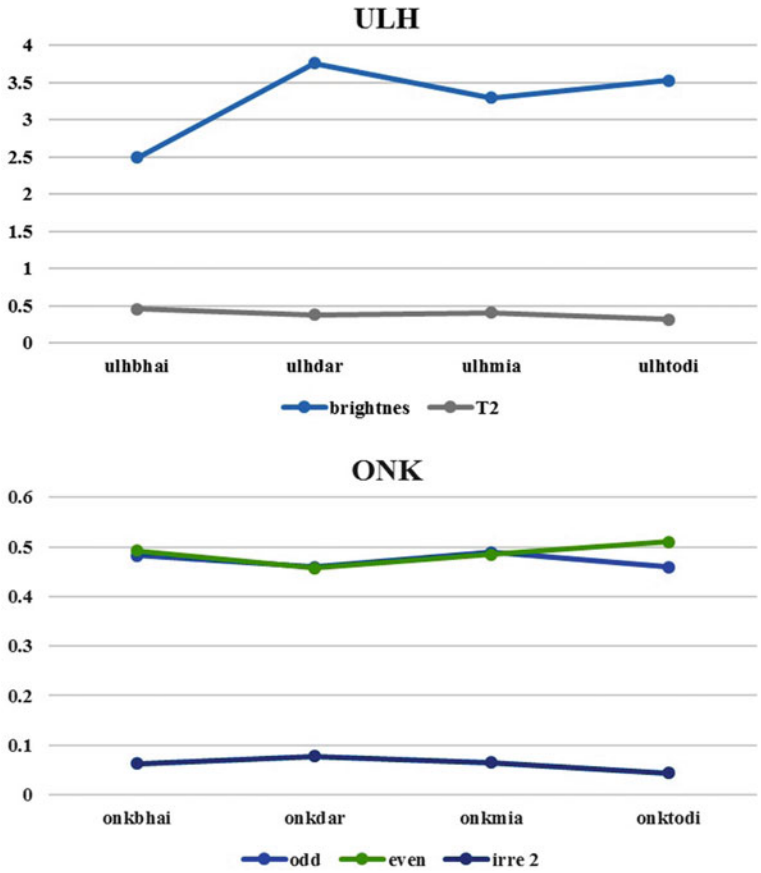


Fig. 1 Similar timbre parameters of the artists ULH and ONK

variations are observed in T1, T2, T3 and spectral irregularity. Uniqueness lies in all the timbre parameters like brightness, spectral centroid, T1, T2, T3, spectral irregularity and odd–even relations for the singers ATK and FAL. The highest degree of similarity in timbre aspects is observed in the signals of ATK and FAL. Uniqueness in amplitude and frequency components is very important cues for style measurement in these two artists.

Timbre similarity is also observed in the renderings of the artist ARN. Parameters like T1, T2, T3 and odd–even relations are similar for this artist. Odd and even partials are very close to each other. The frequency components are similar in all the renderings, and hence almost no variations are observed in T1, T2 and T3. Additionally, little variation in spectral irregularity was observed. Energy levels of all the partials are uniform and higher. Timbre similarity is highly visible in three renderings of the artist ARS, but a small difference is observed in raga Mia-ki-malhar. Uniqueness lies

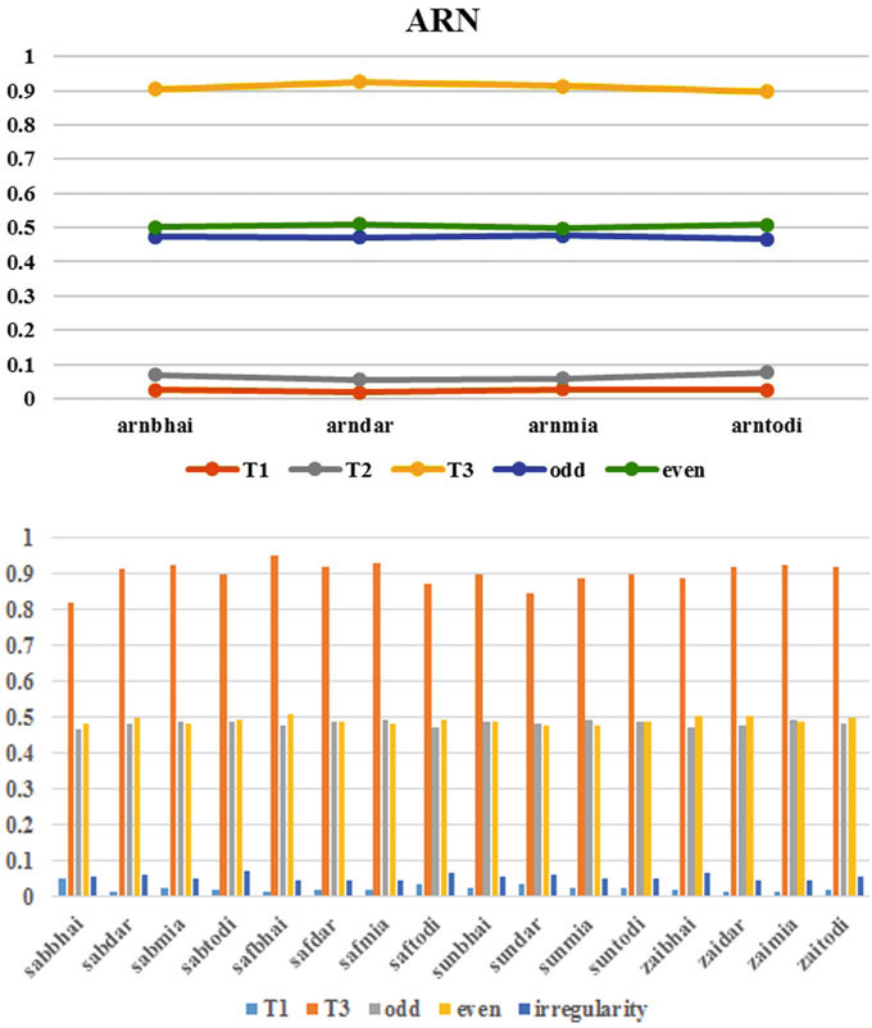


Fig. 2 Similar timbre parameters for the artists ARN, SAB, SAF, SUN and ZAI

in all the timbre parameters like brightness, spectral centroid, T1, T2, T3, spectral irregularity, and odd–even relations, except for the raga raga Mia-ki-malhar.

Very little timbral similarity is observed for the artist ONK. Similarity is only observed in the irregularity parameter. A low timbre similarity indicates a varied voice quality for the artist ONK. Timbre similarity is highly visible in the renderings of the artist SAB, SAF, SUN and ZAI. In these signal, parameters like T1, T3, spectral irregularity and odd–even relations are similar. Odd and even partials are almost overlapping each other. Uniqueness lies in the fundamental and high frequency partials.

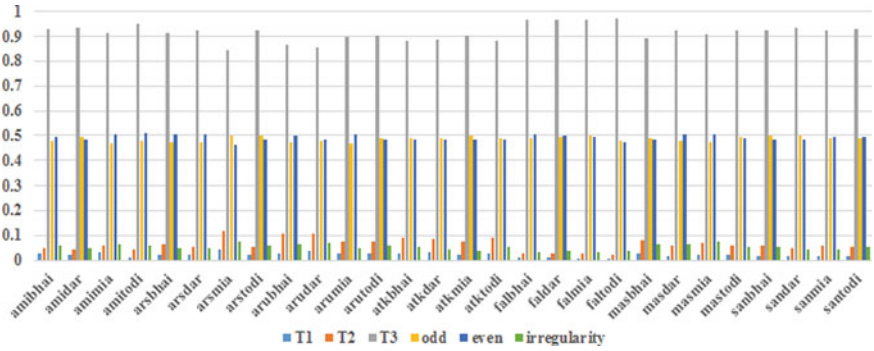


Fig. 3 Similar timbre parameters for the artists AMI, ARS, ARU, ATK, FAL, MAS and SAN

The variations of frequency components are very small in all the renderings, and hence almost no variations are observed in T1, T3 and spectral irregularity.

Brightness is the measure of the centroid of amplitude, i.e. the measure of intensity of the partials in a signal. Uniqueness of brightness is observed in ABD, ATK, ULH and FAL. Therefore, maximum timbre quality and similarity in timbre parameters are observed in the artists ATK and FAL.

The spectral centroid is the measure of center of gravity of the frequency components in a signal [9] and it may correlate with the subjective quality of brightness [10]. The centroid is also the measure of the central tendency of frequency of the signal's partials. Uniqueness in the spectral centroid of four signals is observed in artist SAF only. This spectral characteristic is an important timbre cue in identifying the style of an artist.

Since odd and even partials are mostly similar for all the artists, this parameter is the least helpful in identifying the style of an artist. Odd–even harmonics is a measure for the energy distribution on even and odd harmonics and is related to the subjective sensation of fullness of a sound. Here, we have observed that most of the signals have similar energy distribution on even and odd harmonic partials. Hence, no instance of nasality and hollowness is observed, because nasality and hollowness in sound is caused by the dominance of odd harmonics.

The tristimulus is an important descriptor of the spectral energy distribution. It measures the energy of the fundamental (T1), the first three partials (T2) and the higher partials (T3) in relation to the integral energy. The spectral energy levels of all partials are found to be unique for eight artists, viz. ARU, ARS, AMI, ATK, FAL, MAS, SAN and ARN. So the tristimulus parameter is an important cue for style identification of a vocalist. The relative rates of energy build-up in the harmonic partials are a salient feature of the voice quality of these artists.

Irregularity is the measure of smoothness of a spectrum. The variation of spectral energy among partial to partial is very small with the twelve artists ARU, ARS, AMI, ATK, FAL, MAS, SAN, ONK, SAB, SAF, SUN and ZAI.

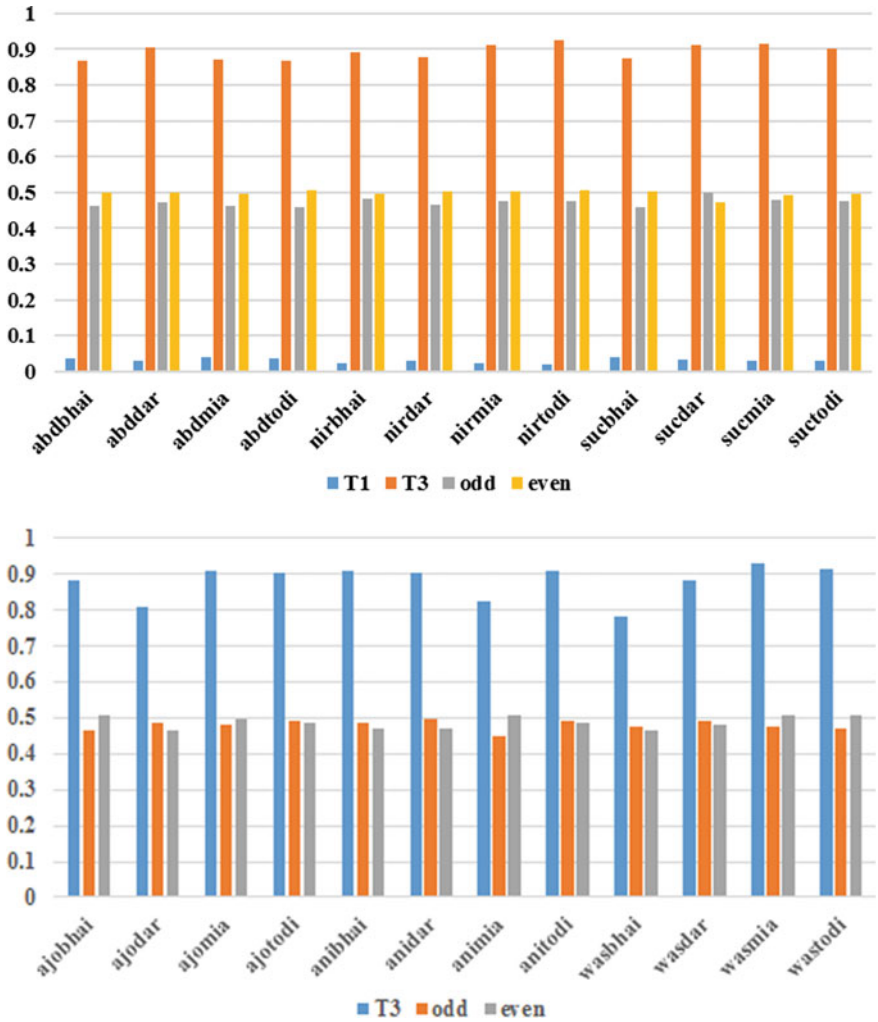


Fig. 4 Similar timbre parameters for the artists ABD, NIR, SUC, AJO, ANI and WAS

Jitter and shimmer are the measures of random perturbations of pitch and amplitude, which are always associated with a quasiperiodic musical wave. Jitter and shimmer are acoustic characteristics of not only the voice signals, but also the instrumental signals. They are measured as the cycle-to-cycle variations of the fundamental frequency; this is called jitter. The measure of peak-to-peak variation of amplitude of the waveform is called shimmer. The measure of the noise on the amplitude envelope is also called shimmer, and the noise on the frequency is also called jitter [12]. Table 1 shows the relative jitter and shimmer of all the sound signals.

The perturbations in pitch and amplitude are very similar in all the four signals of the artist FAL. The vocal characteristics are unique for this artist. This is a striking

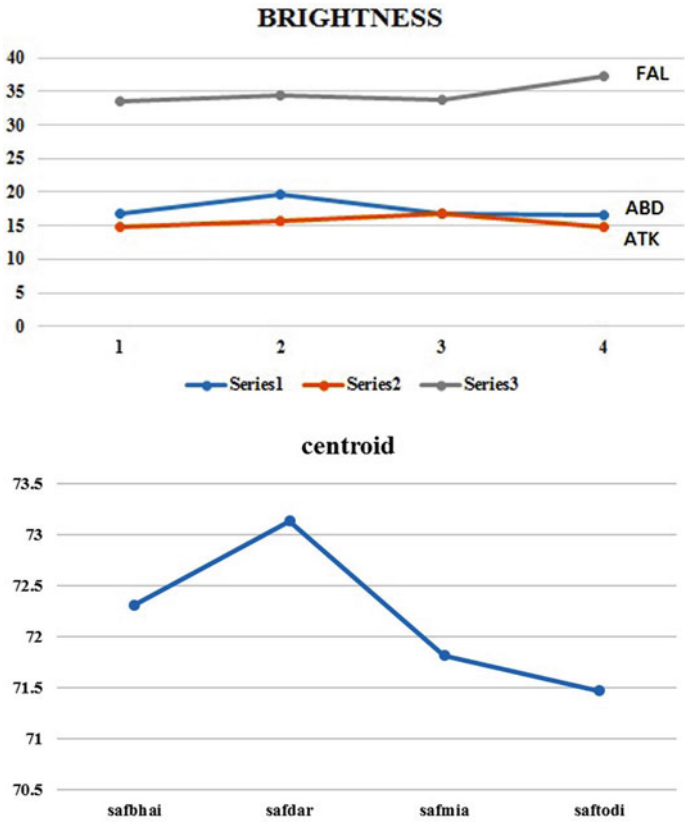


Fig. 5 Similarity found in brightness and centroid

feature of this artist and definitely an important cue for style analysis. Another interesting observation is obtained in all the four signals of the artist ULH. Both the perturbations in pitch and amplitude in all the four signals of the artist ULH are very high compared to the other artists, and hence the voice quality is completely different from the other artists. This may be due to the breadth noise. Jitter and shimmer are very important cues for identifying the style of vocal performance of this artist. Similarity in perturbations in pitch and amplitude is also observed in all the four signals of the artist SAN. The vocal characteristics are unique for all four signals of this artist. This is a striking feature of this artist and definitely an important cue for style analysis. It is also observed that the perturbations in pitch are only similar in all four signals of the artists ATK and SUC. Perturbations in amplitude are only similar in all four signals of the artists SUN and ZAI. No significant similarity in perturbation in pitch and amplitude is observed, and hence no uniqueness in voice quality is observed in the sound signals of the other artists, except ATK, FAL, SAN, SUC, SUN, ZAI and ULH. Another interesting fact is that the (relative) value of jitter is not significantly different among the artists, but the (relative) value of

Table 1 Jitter (relative) and shimmer (relative) of all signals

Jitter					Shimmer				
	Bhai	Dar	Mia	Todi		Bhai	Dar	Mia	Todi
ABD	6.478	4.6806	6.494	9.1967	ABD	−0.1319	0.0616	−0.5742	0.6907
AJO	8.4725	19.0123	5.6649	5.7759	AJO	1.8532	2.381	0.746	1.0594
AMI	**5.0264**	**4.6722**	**6.294**	**4.4778**	AMI	0.8671	0.6957	0.8184	2.3558
ANI	5.824	6.6683	17.5246	6.2409	ANI	−0.8538	0.3287	3.2026	1.2623
ARN	4.8641	5.4866	4.5842	7.0639	ARN	0.2286	1.8913	0.2797	1.8396
ARS	5.5476	5.7811	11.2242	5.6368	ARS	1.0818	1.4055	1.9243	1.6062
ARU	9.8627	9.7326	6.9487	7.5292	ARU	2.4753	1.9281	0.9975	1.2604
ATK	*5.9514*	*6.7568*	*5.9009*	*6.6934*	ATK	−0.1339	0.1876	1.0126	0.8312
FAL	*3.6921*	*3.3343*	*3.3666*	*3.1812*	FAL	*1.4004*	*1.3503*	*1.4851*	*1.3463*
MAS	7.6111	6.8994	9.2211	7.1041	MAS	1.9643	2.0352	2.1281	1.2959
NIR	8.9793	9.7345	6.0261	6.0681	NIR	2.628	1.6815	1.455	1.9174
ONK	7.435	16.8036	9.989	5.6202	ONK	1.9479	−0.915	2.3179	−0.253
SAB	8.453	7.9083	4.5828	7.894	SAB	0.021	2.7307	0.1173	2.3578
SAF	3.9884	5.421	4.6379	7.4288	SAF	1.4377	0.9417	0.8611	0.311
SAN	*5.7188*	*4.0141*	*4.7511*	*5.3145*	SAN	*1.1864*	*0.8521*	*0.8783*	*1.2288*
SUC	*5.6477*	*6.3761*	*5.143*	*5.7903*	SUC	−0.5455	0.5935	0.3571	0.3855
SUN	6.4506	8.6375	5.6942	5.928	SUN	*1.4591*	*0.8951*	*1.0048*	*1.0249*
ULH	***37.037***	***26.4865***	***33.6***	***36.8182***	ULH	***−34.8207***	***−8.899***	***−11.1702***	***5.1702***
WAS	8.4666	7.35	5.5796	5.4883	WAS	−0.4655	1.3853	1.8172	0.856
ZAI	9.4947	5.6102	5.7077	7.6678	ZAI	*2.4241*	*1.5439*	*1.7286*	*1.8824*

shimmer is significantly different among the artists. That means that the perturbation in pitch is almost same among all the artists, but a significant difference is observed in the perturbation in amplitude among all the artists. Therefore, this voice quality measurement may be focused on the singer identification system. It was reported that significant differences can occur in jitter and shimmer measurements between different singing styles, especially in shimmer measurement.

Spectral irregularity is the measure of spectral smoothness [11]. Irregularity among partial shows the measure of irregularity of a given partial with respect to its preceding and post-partials in a signal. It is computed with the short-term Fourier transform where the mean of the given, next and previous amplitude values are compared with the given amplitude value. Bregman remarked that the smoothness of a spectrum is an indicator for partials belonging to a same sound source, and a single higher intensity partial is more likely to be perceived as an independent sound [6]. It has also been found to be useful in revealing complex resonant structures of a sound source. Standard deviations of irregularities among partials were measured for each signal and these standard deviations are compared with four signals of each artist, as shown in Fig. 6.

From this graph, a higher level of spectral smoothness is observed in the signals of the artist ABD, AMI, ANI, ARN, ARS and ARU. This means that the spectral envelopes of vocal sound signals of these artists are almost continuous. Again, spectral roughness is observed in the signals of the artists ATK, AJO, FAL, MAS, NIR, ONK, SAB, SAF, SAN, SUC, SUN, ULH, WAS and ZAI. In principle, the vocal systems separate wider from one signal to the other, and spectral envelopes of their sounds may be assumed to be non-continuous.

Deviation of the irregularity of partials from the mean is much less for the sound signals of the artists ABD, AMI, ANI, ARN, ARS and ARU. It is also observed that such deviation of the irregularity of partials from the mean value is almost similar for all four signals of each of these artists. The spectral energy is uniform in all four sound signals for these six artists. Minimal irregularity in amplitude among partials

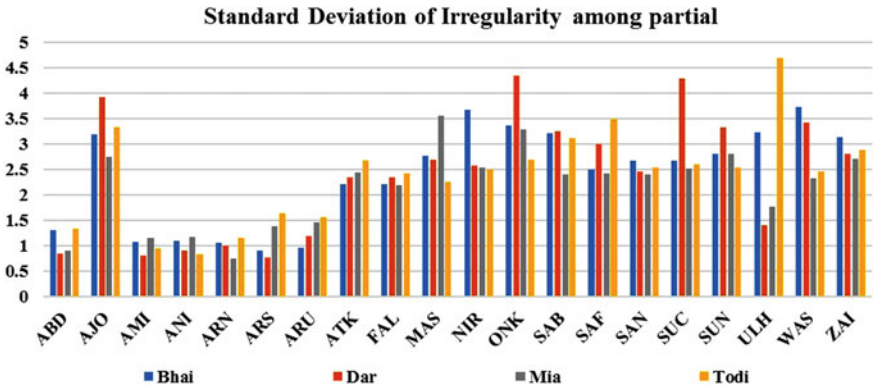


Fig. 6 Standard deviation (along *y*-axis) of irregularity among partial from its pre- and post-partials

is an important voice quality characteristic for these six artists and an important cue for style identification. On the other hand, deviation of irregularity of partials from the mean is higher for the sound signals of the artists ATK, FAL and SAN. It is also observed that such deviation of irregularity of partials from the mean value is almost the same for all four signals of each of these artists. The spectral energy is uniform in all four sound signals for these three artists. A higher but uniform irregularity in amplitude among partials is an important voice quality characteristic for these three artists and an important cue for style identification. For the artists other than these nine, it is observed that the irregularity among partials is higher, but no similarity is observed among the four signals of each artist. The spectral energies among the sound signals of these artists are not uniform.

Smoothness in frequency is measured by frequency per partial number. Standard deviation of frequency per partial number is shown in Fig. 7. Deviation of frequencies among partials from the mean frequency is almost similar among the four sound signals of the artists ABD, MAS, ATK, SAN, SUN, FAL and ZAI. Frequency distribution in the spectral envelopes of all the four signals of these five artists is uniform, and this is an important feature in identifying the style of a vocalist. Again, deviation of frequencies among partials from the mean frequency varied a little among the four sound signals of the artists AJO, AMI, ANI, ARN, NIR and SAF. Frequency distribution in the spectral envelope of all the four signals of these six artists is less uniform. Smoothness in frequency distribution among signals is one of the unique characteristics of the sound signal.

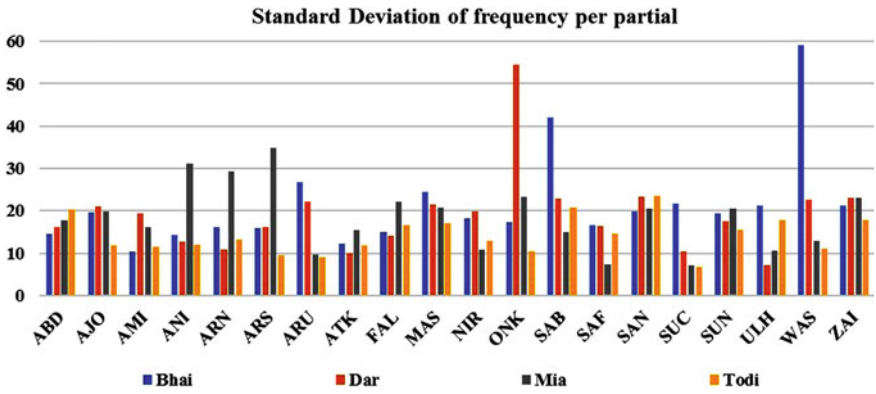


Fig. 7 Standard deviation (along *y*-axis) of frequency per partial number

3 Conclusion

A very high timbre similarity is observed in the signals of FAL and ATK, while very low timbre similarity is observed in the signals of the artist ONK. Energy of the fundamental, mid-frequency partials and high frequency partials is quite high in the signals of the artists ARU, ARS, AMI, ATK, FAL, SAN, ARN and MAS. In the sound signals of ABD, NIR, SUC, AJO, WAS, ANI, SAB, SAF, SUN and ZAI, it has been observed that the energy is being pumped up in the high frequency partials. A unique spectral centroid is observed in the signal of SAF. Odd and even partials are mostly similar for all the artists. From a spectral point of view, a higher level of spectral smoothness is observed in the signals of the artists ABD, AMI, ANI, ARN, ARS and ARU. Again, spectral roughness is observed in the signals of the artists ATK, AJO, FAL, MAS, NIR, ONK, SAB, SAF, SAN, SUC, SUN, ULH, WAS and ZAI. Uniform spectral energy is observed in the sound signals of the artists ABD, AMI, ANI, ARN, ARS and ARU. A higher but uniform irregularity in amplitude among partials is an important voice quality characteristic for the artists ATK, FAL and SAN. Frequency distribution in the spectral envelope of all four signals of the artists ABD, MAS, ATK, SAN, SUN, FAL and ZAI are similar. Perturbations in pitch and amplitude are quite similar in all four signals of the artist FAL. The vocal characteristics are unique for the artist FAL. Perturbations in pitch and amplitude in all the four signals of the artist ULH are very high compared to the other artists due to the existence of the breadth noise, and hence the voice quality is completely different from the other artists. Therefore, timbre is an important feature to distinguish or identify the style of a vocalist.

References

1. Seema G, Shivam S, Kumar MV (2017) Acoustic analysis of Indian classical music using signal processing methods. IEEE, Penang, Malaysia, ISSN: 2159-3450, TENCON 2017
2. Datta AK, Solanki SS, Sengupta R, Chakraborty S, Mahato K, Patranabis A (2017) Signal analysis of Hindustani classical music. Springer Nature. ISBN 978-981-10-3959-1 (eBook)
3. Jensen K (2002) Perceptual and physical aspects of musical sounds. University of Copenhagen, Universitetsparken1, Copenhagen, Denmark
4. Pollard HF, Jansson EV (1982) A tristimulus method for the specification of musical timbre. Acustica 51:162–171
5. Jensen K, Georgios M (2001) Hybrid perception: 1st seminar on auditory models, Lyngby, Datalogisk Institute, University of Copenhagen, Universitetsparken1. Denmark, Copenhagen
6. Bregman A (1990) Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge
7. Hong PT (2004) Towards automatic musical instrument timbre recognition. Ph.D. thesis, Department of Music, Princeton University
8. Joao PT, Andre G, Teixeira JP, Goncalves A (2016) Algorithm for jitter and shimmer measurement in pathologic voices. Science Direct
9. Beauchamp J (1982) Synthesis by spectral amplitude and Brightness matching of analyzed musical instrument tones. J Acoust Eng Soc 30(6):1982

10. McAdams S, Winsberg S, Donnadieu S, de Soete G, Krimphoff J (1995) Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol Res 58:177–192
11. McAdams S, Beauchamp JW, Meneguzzi S (1999) Discrimination of musical instruments sounds resynthesized with simplified spectrotemporal parameters. JASA 104(2)
12. Richard G, Richard C, d'Allesandro (1996) Analysis, synthesis and modification of the speech aperiodic component. Speech Commun 19

Style Identification of Vocal Singers in Indian Classical Music Using Meend and Andolan

Kaushik Banerjee, Anirban Patranabis, Ranjan Sengupta, and Dipak Ghosh

Abstract In Hindustani music, a small segment is the precomposed but the remains involve with a gradual systematic exploration of the raga by various forms of improvisation keeping the grammar of the raga intact. Performers are free to explore their improvisational patterns based on their experience and expertise. According to the artist's cognitive and imaginative abilities, the quality and nature of improvisation differ among artists. One of the most important tools used in improvisation is musical ornamentation (alankar/alankaran). Extempore variations that a musical performer creates while performing within the limit of melodic pattern and the rhythmic cycle could be termed as alankar or alankaran (ornamentation). These variations in performance embellished and enhanced the beauty of the raga. Here, in this study, we are interested in finding the hidden musical patterns and structures in Indian classical music which are fundamental in unfolding the musical ornamentation of a raga. Here, we have discussed only two ornamentations, viz. meend and andolan. The study leads to identifying the style of vocal performers.

Keywords Hindusthani music · Ornamentation · Meend · Andolan

1 Introduction

Indian classical music has traditionally been developed about a single-line melodic pattern, but artists do very complex things with their experience, expertise and the mood of their performance with that one line of melody. Simple melodies are miraculously brought to life through ornamentation (alankar) [1]. So, what exactly is ornamentation? The assumption is that there is an underlying skeleton of melody that can stand on its own; ornamentation is what is added to this to make it more appealing. Ornamentation can take place at the level of individual notes—when a note is sung in any manner except straight, it is said to be ornamented. Some ornaments are applied to groups of notes. There can even be ornamentation at the level of the composition

K. Banerjee (✉) · A. Patranabis · R. Sengupta · D. Ghosh
Sir C V Raman Centre for Physics and Music, Jadavpur University, Kolkata, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_7

itself, such as by introducing melodic or rhythmic variations. Ornamentation or decorations of a musical performance is an indispensable part of improvisation in Indian classical music, which is not allowed in western classical music [2]. The performer tries to evoke the mood or essence of the raga (melodic-mode) with the delicate exploration by flourishing the ornamentation. Unlike western classical music, Indian classical music is cognitively associated with these ornamentations instead of the actual underlying notes only [3]. There are many different kinds of ornaments. Some add finer nuances to the melody, and others give it a new dimension in the form of texture. Together, the various ornaments play a very important role in giving body and expressiveness to a simple melody, making it complete in and of itself without the need for accompaniment.

The objective of this study is to computationally detect the ornaments in Khayal performances by various eminent vocal singers in Indian classical music. In HM, we have the most important form of annotations of primary alankaras (ornamentations), namely meend (glides between notes), andolan (oscillations in a note region), gamaka (two notes alternation), murki (oscillation among two notes touching the adjacent note) and khatka (singing a note, touching a higher or lower note, before landing on the main note). This paper makes an attempt to study and discuss only two of them, viz. meend and andolan computationally and attempt to identify style of the artists based on these two ornamentations. For this, eighty-two songs sung by 21 singers covering four ragas namely Bhairav, Todi, Darbari Kannada and Mian-Ki-Malhar were taken for analysis. Only the aalap (only the vocal part of a rendition without any influence of accompanying instruments like table, etc.) portion of various ragas sung by various eminent singers in Indian classical music was considered. Minimum duration and maximum duration of the signals are in between 3 and 6 s. Most of the singers are eminent and well known in Indian classical music. Digitization of the signal was done @44,100 samples/s (16 bit/sample).

2 Meend

The meend in its most basic form can range from a simple span of two notes to a whole octave. These are straightforward, smooth and unidirectional. The basic meend is generally very slow paces and usually rendered in the first part of the alaap-vistaar. As the pace gradually picks up, the meends also gain in tempo and progress to more complex structures (www.itcsra.org). The role of transitions is also relevant to the concept of ornamentation in Indian music which is essential to embellish or enhance the inherent beauty of the genre [4]. We have used the algorithm by Datta et al. [5–8] to extract notes from each signal and procedure for the extraction of meend from the audio signal. Figure 1 shows the nature of six different transition categories. In this, the 1st figure shows the pattern of rising while the second figure shows the fall. The majority of these two types of meend comprise with two to four note transitions. The third figure shows the hat category of transition. In this, we found two subcategories of hat, viz. erect hat and inverted hat but we have measured

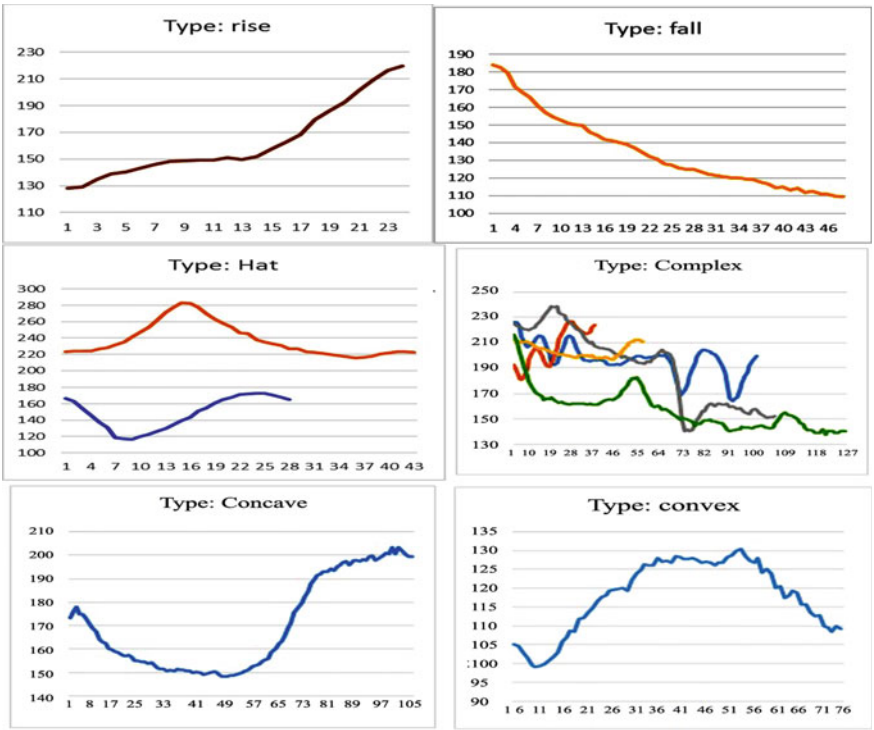


Fig. 1 Different categories of transition pattern

the duration and frequency of occurrences of both the category as a whole in a single category 'hat'. In this category, the artist glides among three to five notes in which one is at the centre and others are at two adjacent sides of it. In the next figure, we showed only a few among various types of the complex pattern of transitions (meend) found in all the renditions. In these kinds of meend, various sets of note combinations were used by the artist ranging from four note combinations to full octave. The last two figures describe how the concave and convex category of transition looks like. These are smoothly gliding patterns of meend. These types comprise of three to four notes.

We have identified the transition among notes of each signal and then we categorize them into six different types. Table 1 shows the frequency of occurrences of transition among notes, which is the measure of number of transitions per second. It is found that all the six categories of transition were found only in the signals of the artist FAL and ULH. Artist FAL showed preferences over a complex category of transition. They used other categories of meend judiciously as per the mood of the raga. Meend types are an important style stamp of the artist FAL. In the signals of the artist, ABD, an interesting observation, was found that the frequency of occurrences of both the concave and convex type of transition among notes were the same. This is a unique

Table 1 Frequency of occurrences of different category of meends

	bhairav						darbari kannada						mia-ki-malhar						todi					
	conc	conv	rise	fall	hat	complex	conc	conv	rise	fall	hat	complex	conc	conv	rise	fall	hat	complex	conc	conv	rise	fall	hat	complex
abd	10	11	33	9	37	0	17	17	25	8	33	0	10	8	10	10	43	14	11	11	14	14	45	5
ajo	80	0	2	6	12	0	62	0	4	11	15	8	46	0	18	9	13	9	44	0	6	21	23	6
ami	14	0	0	57	29	0	26	0	26	11	37	0	18	0	14	45	23	0	18	0	14	32	27	9
ani	40	0	60	0	0	0	26	0	21	11	32	11	0	10	30	40	20	0	13	0	38	25	25	0
arn	14	7	7	29	36	7	10	10	23	29	29	0	15	11	11	11	37	15	0	0	15	8	23	54
ars	0	0	19	19	38	24	0	0	15	16	46	23	0	0	21	11	47	21	0	0	28	31	25	16
aru	0	0	48	24	29	0	7	7	47	13	20	7	0	0	26	16	37	21	0	0	36	27	27	10
dol	0	0	31	15	31	23							0	0	18	29	24	29	0	0	15	23	46	16
fal	20	0	13	13	13	40	5	5	27	36	14	14	0	4	30	22	9	35	4	0	24	28	8	36
mas	6	6	38	13	31	6	28	0	20	20	28	4	7	0	27	33	27	7	6	0	18	29	12	35
nir	0	0	27	36	27	9	0	0	25	13	63	0	0	0	44	17	22	17	0	0	38	38	25	0
onk	0	0	10	20	40	30	0	0	21	21	50	7	0	0	27	18	36	18	0	0	8	35	42	15
rup	0	0	27	45	27	0	0	0	0	0	0	0	7	0	29	7	50	7	0	0	13	25	31	31
sab	0	0	20	20	13	47	5	0	15	35	40	5	0	0	26	32	16	26	5	0	15	40	25	15
saf	0	0	19	13	44	25	11	6	17	11	22	33	6	3	15	12	33	30	0	0	28	31	41	0
san	0	6	24	29	24	18	5	0	20	45	15	15	0	0	23	23	31	23	0	0	29	7	29	36
suc	5	0	21	26	32	16	0	0	21	21	0	57	0	0	25	20	40	15	3	5	24	26	24	18
sun	0	0	33	33	33	0	0	12	36	33	15	3	0	10	24	19	29	19	7	0	13	20	33	27
ulh	7	0	13	20	33	27	3	3	21	31	34	7	4	0	12	42	38	4	0	0	20	37	23	20
was	10	0	19	24	33	14	0	0	7	27	27	40	0	0	16	26	26	32	8	0	13	33	33	13
zai	0	0	62	15	15	8	0	0	24	35	12	29	10	0	30	30	0	30	5	0	36	36	9	14

style of this artist, and such observations were not found in the signals of other artists. Majority of transitions are found to be hat type in the signals of the artist ABD. Other categories of meend were used judiciously as per the mood of the raga. Majority of transition found in the signals of the artist AJO is concave type. No convex type of transition is found in the signals of this artist. He showed his preferences over the concave category of meend and this may be a style stamp for this artist. Again no convex category of meend was found in the raga renderings of the artist WAS. Other categories of transition among notes were used as per the then mood of the particular raga and hence no specific pattern was found to be unique. No specific observations in different categories of transitions were found in the sound signals of the artists AMI, ARN, ARU, SAF, SAN and ANI and hence no style stamp can be drawn in terms of transition among notes for these artists. No concave and convex types of transition among notes were found in the renderings of the artists ARS, ONK, RUP, SUC and DOL. All meends are limited to the hat, rise, fall and complex type. But in the renderings of the artist SUC, the frequency of occurrences of the complex category of transition among notes was found to be high. In the signals of the artist MAS, no convex type of transition among notes was found. But the frequency of the concave and complex type of transition was too low. Most of the meend types are limited to rise, fall and hat category and this is a style stamp of this artist. No concave, convex and complex category of transition among notes were found in the renderings of the artist NIR. Meends of this artist are limited to rise, fall and hat category only. No concave and convex types of transition among notes were found in the renderings of the artists SAB, SUN and SUC. Interesting observation found in their renderings is that the frequency of occurrences of other four types of meends is almost similar among four ragas. Complex category of meend is quite high in the signals of the artists SUN and SUC. In his renderings of the artist ZAI, no concave- and convex-type transitions were found. Mostly rise category followed by the fall category are two mostly utilized transition type found in his renderings. If we consider the overall result, we found that the maximum occurrences of meend are rise, fall and hat category while least are of convex category.

Now minimum duration and maximum duration of transition were identified for each of the raga renderings. The average duration of transition and frequency of occurrences of transition were also measured and are shown in Table 2. Minimum, maximum and average duration of the transition and also the frequency of occurrences of transition are unique and identical in the renderings of the artists AJO, FAL and DOL. Although the notes are different and the type of transitions are different, these four parameters show uniqueness and hence transition among notes is definitely a style stamp of these three artists. The average duration of transition for the artist AJO is 2.21 s, for the artist FAL, it is 3.08 s, and for the artist DOL, it is 2.56 s. The long durational transition among notes is rarely observed in all the renderings of the artist AJO and DOL. But the artist FAL sung long durational transition among notes very often and a wide range of duration of transitions are observed in his renderings. The average frequency of occurrences of meend for the artist DOL is 2.84, that for AJO is 3.58 and 5.51 for FAL.

Table 2 Minimum, maximum and average meend duration (sec) and frequency (%) of occurrences of meend

	Bhai				dar				mia				todi			
	min	max	mean	freq	min	max	mean	freq	min	max	mean	freq	min	max	mean	freq
abd	1.28	3.12	2.2	3.67	1.24	5.57	2.9	2.33	1.45	8.63	3.3	5	0.76	7.58	3.44	4.33
ajo	1.13	3.06	2.52	3.33	1.92	3.95	2.96	3.33	1.28	3.2	2.04	4	1.23	3.02	1.96	3.67
ami	0.7	3.13	2.29	5.67	0.41	1.94	1.27	5	0.99	2.3	1.54	6	0.76	10.06	2.08	5.67
ani	1.92	2.79	2.27	1.67	1.02	4.8	2.57	2.33	1.86	9.63	2.88	2.33	1.3	4.54	2.3	2.33
arn	0.67	5.53	2.09	3.33	1.57	3.61	2.27	6.33	1.16	4.11	2.28	6.33	0.64	6.36	2.88	4.67
fal	1.8	4.93	3.04	5	0.97	5.68	2.82	5.33	1.74	5.52	3	5.67	0.92	5.88	3.4	6.03
ars	0.79	3.13	1.91	4.67	0.52	2.15	1.25	3.33	0.85	5.24	2.15	5.33	1	3.95	1.71	7
aru	1.09	4.54	2.24	5	0.81	3.14	1.65	3.67	0.89	5.72	2.31	4.33	1.01	3.27	1.92	5.33
mas	1.38	6.52	3.19	3.67	1.55	4.51	2.5	5	0.97	4.76	2.26	4.33	1.33	6.55	2.62	5
nir	1.22	6.72	3.94	2.67	2.07	5.4	3.31	2	1.11	13.35	4.32	4.33	1.22	6.84	3.3	3
onk	1.02	6.66	2.21	2.33	0.89	3	2.07	2.67	0.56	3.04	1.82	2.33	0.85	5.15	2.05	2.33
sab	2.63	7.63	4.23	4	1.95	6.28	4.09	4	2.49	9	5.26	4.33	1.84	8.06	3.35	4.67
saf	0.57	4.55	2.34	3.67	1.05	3.67	2.13	4.67	0.8	5.15	2.54	7.33	0.58	4.4	1.8	10.67
san	1.69	11.97	5.16	3.67	0.79	11.4	4.53	4.33	3.46	11.87	7.96	3.67	1.35	12.06	6.65	3.67
suc	0.72	8.9	3.09	4.67	2.66	14.31	8.11	3.67	1.3	9.86	5	5	0.52	10.82	4.05	8
sun					1.37	3.42	2.24	1.67	1.29	4.1	2.04	8.33	0.76	3.63	1.64	5.67
ulh	1.08	7.07	3.26	4	1.57	5.96	3.56	5.67	0.86	7.77	3.68	5	1.08	5.6	2.89	7
was	1.82	7.65	3.95	4.33	0.67	10.92	5.25	4	1.38	9.41	5	5.67	0.98	7.41	3.64	5.33
zai	2.4	6.74	4.6	3	2.5	9.57	5.76	4.67	3.09	8.94	5.92	2.67	1.18	7.78	4.12	5.33
dol	1.1	3.63	2.35	3.67					1.2	3.14	3	4	1.21	3.31	2.27	3.67
rup	1.54	8.77	3.6	2.67					1.51	6.99	2.66	2.67	1.65	9.48	4	4

(continued)

Table 2 (continued)

	Bhai				dar				mia				todi			
	min	max	mean	freq	min	max	mean	freq	min	max	mean	freq	min	max	mean	freq
sal	0.25	1.77	0.646	4.95	0.3	1.8	0.659	4.8	0.36	1.66	0.702	3.31	0.33	1.97	0.73	4.62

Minimum and average duration of the transition and also the frequency of occurrences of transition are unique and identical in the renderings of the artists AMI, ANI and ONK. Most of the transition time is low for all three artists. Average transition time for the artist AMI is 1.7 s, for the artist ANI it is 2.7 s and it is 2.2 s for the artist ONK. Only ANI very rarely sung long durational meend and all the three artists preferably sung low durational meend. High frequency of occurrences of meend is observed in the renderings of the artist AMI, and low frequency of occurrences of meend is observed in the renderings of the artist ANI and ONK. Average frequency of occurrences of meend for the artists AMI, ANI and ONK is 5.6, 2.2 and 2.4, respectively. Minimum duration, maximum duration and average duration of transition are unique and identical in the renderings of the artists SAF, SUN and SAL. But frequency of occurrences of transition is different in different signals. Most of the meend durations are low for all the three artists. Average transition time for the artist SAF is 2.14 s, for the artist SUN it is 1.92 s and it is 1.2 s for the artist SAL. All the three artists sang long durational meend occasionally and all the three artists preferably sung low durational meend. Frequency of occurrences of meend is observed to be high for all these three artists. Average frequency of occurrences of meend for the artists SAF, SUN and SAL is 6.6, 3.9 and 4.4, respectively.

No specific similarity and uniqueness are observed in the signals of the artists SUC and ZAI. So style of these artists cannot be identified from transition among notes. Minimum and average duration of transition are unique and identical in the renderings of the artists ARN, ARS, ARU, MAS and ULH. No uniqueness is observed in maximum duration and frequency of occurrences. Occasional long durational meend is observed in the renderings of the artists ARN and MAS while artist ULH preferred to sung more long durational meend. Artists ARS and ARU preferably sang low durational meend, and average duration of transition among notes for the artists ARN, ARS, ARU, MAS and ULH is 2.4 s, 1.8 s, 2 s, 2.8 s and 3.3 s, respectively. Frequency of occurrences of meend is observed to be high for all these five artists. Average frequency of occurrences of meend for the artists ARN, ARS, ARU, MAS and ULH is 5.2, 5.1, 4.6, 4.5 and 5.4, respectively. Maximum and frequency of occurrences of transition are unique and identical in the renderings of the artists SAN while minimum and frequency of occurrences of transition are unique and identical in the renderings of the artists SAB. Most of the transition among notes were found to be long durational for both the artists. Average duration of transition among notes is 4.2 s for the artist SAB and 5.9 s for the artist SAN. Average frequency of occurrences of meend for the artists SAB and SAN is 4.3 and 3.8, respectively.

3 Andolan

An andolan is a slow oscillation applied to a note. It usually features in ragas that use microtones and is applied to the notes in those ragas that involve the use of microtones. Microtones are unstable pitches that fall between two notes and are difficult to sustain. Artists can use this natural instability to their advantage by mastering the

andolan and oscillating the note in a controlled fashion (www.raag-hindustani.com). We have identified all the oscillatory notes and measured the duration of oscillations about a note from each of the signals. About every oscillatory notes, microtones are associated with them. Here, we are interested in main notes not the microtones. So we ignore the microtones. The notes used in andalit swar are raga specific, and the notes are fixed as shown in Table 3. All the artists strictly adhere with the musical grammar of Indian classical music.

The spectrogram of oscillatory notes is in general and looks oscillatory type with formation of crest and trough. Figure 2 shows only two forms of spectral view of andolans of which the first figure is the spectral view of a short durational andolan swara and the second one is that of a long durational andolan swara. Comparing all the signals we have found that the shortest duration of an andolan obtained here is 0.03 s and longest duration of an andolan obtained is 13.983 s. So a wide range of andolan duration enhances the essence of raga. Using oscillatory note in a raga rendering is in general, an inherent property of the artist which he/she expertises through practice and gharana and become an integral part of his/her performance. Again in a raga performance artist very often oscillates a note intentionally as per the mood of the artist and the response from the audiences. Since the notes are fixed for a raga, the duration of oscillatory notes and frequency of occurrences of oscillatory notes is important cue to identify the style of an artist.

We have measured the minimum and maximum duration of oscillations about a note for each raga. Average duration of oscillations about a note and frequency of occurrences of the oscillatory notes (total number of oscillatory notes per second) was also measured. These are shown in Table 4. Uniqueness in minimum and maximum duration of oscillatory notes is observed in all the four raga renderings of the artist ARS. Majority andolan duration is observed to be within 2.5–3 s but no andolan duration less than one second is observed in any signals. Although the oscillatory notes are different for different renderings yet the average duration of oscillatory notes and frequency of occurrences of oscillatory notes are identical. Similar behaviour is shown in all the raga renderings of the artist SAF. All these four categories of oscillatory note durations are similar in all four ragas and majority andolan duration is observed to be within a range of 2–2.5 s. Such uniqueness in duration of oscillatory notes is a unique characteristics of singing style of the artists ARS and SAF. So duration of oscillatory notes is definitely a style stamp of these artists. Oscillatory notes are highly similar in raga Bhairav and Mian-ki-malhar and also they are highly similar in raga Darbari and Todi for ONK. Little differences are observed in these two sets. Uniqueness lies in minimum, maximum and average duration, and also frequency of occurrences of oscillatory notes is observed in each set of ragas of the

Table 3 Oscillatory notes (notes about which the oscillation takes place)

Bhairav	Darbari	Mian-ki-Malhar	Todi
Komal (flat) re	Komal (flat) ga	Ni	Komal (flat) ga
Dha	Komal (flat) dha	Komal (flat) ga	Dha

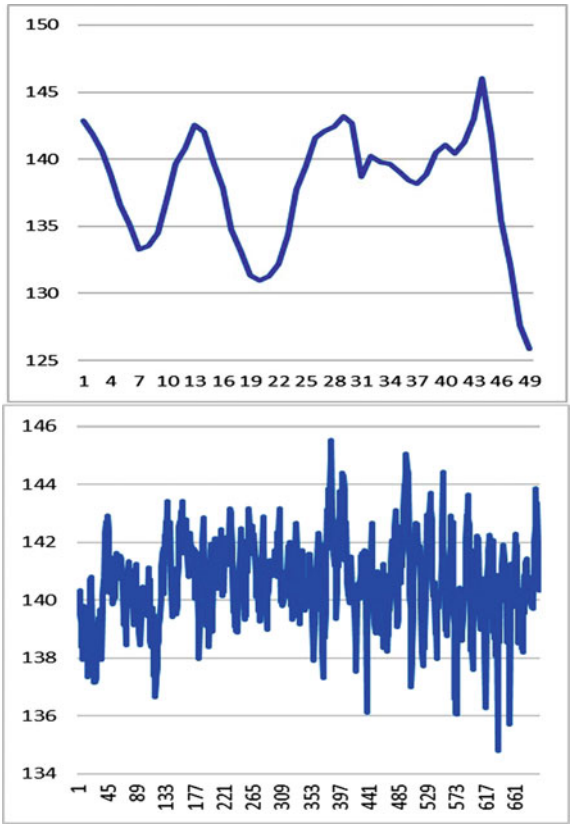


Fig. 2 Shortest and longest andolan type (both are about komal re and from raga bhairav)

artist ONK. No andolan duration less than one second is observed in the renderings of this artist and majority andolan duration is observed to be within 3.5–4 s. No uniqueness is found in the duration of oscillatory notes in the renderings of the artist SUN. But all oscillatory notes sung by the artist are small durational and average duration found to be 1.48 s. This might be a style stamp for this artist. Little differences are found in minimum, maximum, average duration and frequency of occurrences of oscillatory notes in the sound signals of the artists WAS and ZAI. No andolan duration less than one second is observed in the renderings of the artist WAS and average andolan duration is observed to be 3.8 s. No andolan duration less than two second is observed in the renderings of the artist ZAI and average andolan duration is observed to be 4 s.

Also, it is observed that the minimum duration of andolan is identical for all the four raga renderings of the artist ABD. Although the oscillating notes are different for different ragas, the minimum durations are identical. No such identity is observed for maximum duration, and average andolan duration shows very little difference among four ragas. No andolan duration less than one second is observed in the renderings of this artist and majority andolan duration is observed to be within 2.5–3 s. Frequency

Table 4 Minimum, maximum and average andolan duration (sec) and frequency (%) of occurrences of andolan

	Bhairav				Darbari				Mia-ki-malhar				Todi			
	min	max	mean	freq	min	max	mean	freq	min	Max	mean	freq	min	max	mean	freq
abd	1.388	5.069	3.342	3.333	1.452	4.93	2.731	3.667	1.654	2.364	2.012	1.333	1.544	7.762	2.556	3.333
ajo	1.375	5.694	3.551	4	1.534	6.426	3.44	2.667	1.385	4.683	2.605	3.333			4.49	0.333
ami	0.922	3.948	2.498	3	1.32	2.1	1.773	3	1.387	2.71	2.079	1.667	0.831	2.577	1.459	1
ani	2.494	8.996	5.397	2	1.634	4.97	3.804	2.667	3.192	10.118	5.764	1.333	2.097	7.732	4.238	1.667
arn	1.251	1.468	1.36	0.667	1.103	4.122	2.155	5	1.14	4.548	2.088	4.667	2.95	4.16	3.5	1
ars	1.842	4.903	3.307	1.333	1.623	4.715	3.996	1	1.701	4.442	3.307	2.333	1.332	4.432	3.416	1
aru	1.833	4.961	3.103	1.667	2.541	5.527	3.527	2.667	2.639	6.179	3.278	1.667	3.038	4.407	3.723	1.667
dol	2.478	6.965	4.141	1.333					1.307	5.754	3.289	2.333	3.429	13.983	10.252	1
fal	1.04	4.285	2.47	2.667	0.03	3.57	1.839	3	1.145	4.26	2.501	3	2.083	2.083	2.083	2.333
mas	2.201	6.47	3.796	2.333	1.78	3.94	2.901	3.333	1.352	4.832	3.003	3	1.887	4.36	3.124	0.667
nir	2.186	8.196	6.015	1.333	5.218	7.534	6.255	1.667	1.96	5.71	3.809	2.333	3.412	6.503	4.876	2.333
onk	2.385	6.07	3.939	2	3.584	4.36	4.087	1	2.092	6.13	3.531	2.333	3.433	4.172	3.877	1
rup	2.436	13.22	7.858	1.667					4.179	5.725	5.061	2.667	1.867	4.27	2.97	1.333
sab	1.964	6.257	3.825	3	2.712	3.171	3.022	2.333	1.974	3.9	2.672	3.333	1.307	8.337	3.392	2.667
saf	1.052	3.363	2.432	1.333	0.931	3.464	2.774	1.333	1.128	3.581	2.366	1.333	1.333	3.33	2.307	1.667
sal	0.922	4.715	2.079	5.634	1.323	3.822	2.844	4.682	1.722	3.235	2.393	3.55	1.694	4.771	2.632	4.167
san	2.471	8.377	6.146	1.667	1.474	6.01	3.636	3.333	1.844	6.733	4.388	2	3.356	8.465	4.991	1.333
suc	3.784	6.99	5.118	1.333	2.47	10.494	5.162	1.667	2.49	6.787	4.648	2	2.02	3.246	2.633	0.667
sun	1.951	3.001	2.489	2.667	1.03	1.301	1.184	1	0.916	1.947	1.543	1.667	1.315	3.044	2.063	2
ulh	1.127	5.749	2.19	4	1	3.01	1.96	3.333	1.575	4.35	2.64	2.333	1.857	1.857	1.857	0.333
was	1.82	6.96	4.348	3.333	2.463	4.569	3.838	2.667	2.478	5.489	3.377	4	1.882	5.876	3.579	2.667

(continued)

Table 4 (continued)

	Bhairav				Darbari				Mia-ki-malhar				Todi			
	min	max	mean	freq	min	max	mean	freq	min	Max	mean	freq	min	max	mean	freq
zai	2.909	6.044	4.964	2.667	1.994	5.299	3.424	3	2.441	5.172	3.725	2.667	1.37	5.2	3.739	2.333

of occurrence of oscillatory note is also identical for three ragas with an exception in Mian-ki-malhar. Hence, the ornament namely oscillatory note is an important cue to identify the style of the artist ABD. Little difference in frequency of occurrences, minimum, maximum and average duration of oscillatory notes is observed in the renderings of the artist AJO with an exception in raga Todi. In raga Todi only one oscillatory note has been found but some sort of similarity is observed in the ornament like andolan in the remaining three raga renderings. No andolan duration less than one second is observed in the renderings of this artist and majority andolan duration is observed to be within 3–3.5 s. Ornament like oscillatory note may be a style stamp of the artist AJO. Minimum andolan duration and average andolan durations show little difference among four raga renderings of the artist AMI but similarity in maximum andolan duration is observed in three renderings except Bhairav. No similarity in frequency of occurrences of andolan is observed for the artist AMI. So andolan alankaran may not be a style stamp of this artist. No relation in minimum and maximum andolan duration is observed in the renderings of the artist ANI but average andolan duration and frequency of occurrences shows very little difference among four ragas. Interestingly no andolan duration less than two seconds is observed in the renderings of this artist and majority andolan duration is observed to be within 4–5 s. Artist ANI prefers to sing long durational oscillatory notes which is definitely a style stamp of his voice.

Identical minimum duration in the oscillatory notes is observed in the renderings except for Todi of artist ARN. Also, identical maximum duration in the oscillatory notes is observed in the renderings except Bhairav. But no similarity is observed in average duration and frequency of occurrences of oscillatory notes in the renderings of the artist ARN. No andolan duration less than one second is observed in the renderings of this artist and also no long durational oscillatory notes are observed in the renderings of this artist. Majority andolan duration is observed to be within 1.5–2 s only. Little difference is observed in minimum and maximum duration of oscillatory notes in the sound signals of the artist ARU. Uniqueness in average duration and frequency of occurrences of oscillatory notes is observed in all the four raga renderings of the artist ARU. No andolan duration less than one second is observed in the renderings of this artist and majority duration of oscillatory notes are observed to be within 3.5–4 s. Uniqueness in average duration of oscillatory notes is observed in the sound signals of the artist FAL but no such uniqueness is observed in minimum and maximum duration of oscillatory notes. Little difference is observed in frequency of occurrences of oscillatory notes in the sound signals of the artist FAL. Uniqueness in maximum duration and average duration of oscillatory notes is observed in the sound signals of the artist MAS. No similarity is observed in minimum duration of oscillatory notes and frequency of occurrences of oscillatory notes are observed in the sound signal of this artist. No andolan duration less than one second is observed in the renderings of this artist and majority duration of oscillatory notes are observed to be within 3–3.5 s. Uniqueness lies in the renderings of the artist NIR is that most of all oscillatory notes are long durational with majority ranging above 5 s although there is no similarity found in minimum, maximum and average duration of oscillatory notes. Little difference is observed in the frequency of occurrences of

oscillatory notes for this artist. Long durational oscillatory note is a style stamp of this artist.

Little difference in minimum duration and frequency of occurrences is observed in the renderings of the artist SAB. Uniqueness lies in average duration of oscillatory notes. This artist used low duration to long duration of oscillatory notes but no duration less than one second is found. No specific pattern of oscillatory notes is found in the signal of the artist SAN. Only information about style in terms of andolan can be obtained from the renderings of the artist SAN is that the artist preferred only long durational oscillatory notes and majority of oscillatory note duration is 4–5 s. Little difference in frequency of occurrences, minimum, maximum and average duration of oscillatory notes is observed in all the raga renderings of the artist SAL. Oscillation pattern is unique in all notes. So this might be a style stamp of this artist. Minimum duration of oscillatory notes is unique for raga Bhairav, Darbari and Mian-ki-malhar while there is no similarity found in maximum duration of oscillatory notes for the artist SUC. Little differences are found in average duration and frequency of occurrences of oscillatory notes. No oscillatory notes were found less than two seconds and this artist prefers long durational oscillations with average duration of 4.7 s. This could be a style stamp for SUC. Little differences are observed in minimum and average duration of oscillatory notes of the artist ULH. No relations are found in maximum duration and frequency of occurrences of oscillatory notes of this artist. No oscillatory notes were found less than one seconds and majority duration oscillations is 2.2 s.

4 Conclusion

In simple terms, meend can be described as a mode of joining two notes by means of a graceful glide. This mode of reaching from one note to another is naturally very artistic as it has the smoothness and grace of gliding or skating instead of any kind of jerks. But the very important advantage of this particular embellishment is that in such a process, all the intermediate notes and even the microtones or Shrutis are slightly touched and the tonal result is an amazing fullness and an endless continuity. In the process of meend, two or more number of notes are involved in general. The meend begins from the first note and continue up to the second note encompassing the frequencies of the in between notes. The pitch difference between these two notes could be shorter or longer.

Artist ABD is the only artist who used both the concave and convex type of meend equally and showed maximum interest over hat category of meend. This is definitely a style stamp. Artist AJO showed maximum interest over concave category of meend only. All six categories of meend were used by the artist FAL and ULH. In the signals of the artists NIR and MAS meend categories are mostly limited to rise, fall and hat category. Artist ZAI showed his preferences over rise and fall category of transition among notes.

High frequency of occurrences and frequent use of long durational meend are important features of style of the artist FAL. Low frequency of occurrences and use of short durational meend are important features of style of the artists AJO and DOL. Minimum, maximum and average duration of transition and also the frequency of occurrences of transition are style stamp of these three artists. Also, long durational transition among notes is observed in the renderings of SAB and SAF. Also, uniqueness in frequency of occurrences and long meend is definitely style stamp of the artists SAB and SAF.

In the context of Indian classical music, the application of note transition plays an important role in embellishing inherent beauty of the genre. It is one of the important style stamp of a vocalist. This style characteristic is well understood by different intonation pattern like flat, rise, fall, hat, valley, etc. In Hindustani music, artists have enough liberty to improvise their performance within constraint of fixed framework of a given raga. Ornamentation or alankar is a part and parcel of improvisation in HM. Proper and useful utilization of alankar bring appreciation from listeners. Simple flat meend with 2 or 3 note combinations or complex type of meend with more notes combination evokes different emotion to the listeners. The type of meend is a unique identification of style of a performer. Some prefer short but repeated use of meend while some prefers long durational slow tempo meend.

Since all the artists are experts in Hindustani classical music, they have mastered in oscillating the note in a controlled fashion. Although the note to be oscillated is raga specific but some similarity has been found in terms of average duration of oscillation and frequency of occurrences of oscillatory note for a given artist. Uniqueness of artist SUN is that the artist sung only short durational oscillatory notes in all his renderings. Also, artists like AMI, ULH, FAL and ARN also sung short durational (average duration found to be less than 1.5 s) oscillatory notes. They hardly sung long durational oscillations. Again artists like NIR, SAN, ANI, DOL, SUC, ZAI and RUP sung long durational (average duration found to be more than 4.5 s) oscillatory notes. They hardly sung short durational oscillations. Average duration of oscillatory note may be an ornamental parameter of a vocal artist. We cannot distinguish the style of an artist based on frequency of oscillation as number of oscillatory notes sung per second lies in between 1.2 per second and 2.9 per second.

Duration and frequency of occurrences of oscillatory notes are unique characteristics of singing style of the artists ARS and SAF and hence ornament like oscillatory note is definitely a style stamp of these two artist. High degree of similarity in oscillatory notes are observed in the renderings of these two artists. These parameters are also style stamp of artists like ONK, WAS, ZAI, ARU and SAL. Although little difference in duration and frequency of occurrences of oscillatory notes is observed in the renderings of these artists, they have a unique style in production of oscillatory notes.

References

1. Rao P, Ross JC, Ganguli KK, Pandit V, Ishwar V, Bellur A, Murthy H (2014) Classification of melodic motifs in raga music with time-series matching. J New Music Res 43(1)
2. Ross JC, Vinutha TP, Rao P (2012) Detecting melodic motifs from audio for Hindustani classical music. In: 13th international society for music information retrieval conference (ISMIR 2012)
3. Ganguli KK, Rao P (2015) Discrimination of melodic patterns in Indian classical music. In: Proceedings of national conference on communications (NCC) 2015, IIT Bombay, India, 27 Feb–1 Mar 2015
4. Sengupta R, Dey N, Nag D (2006) Extraction and relevance of transitory pitch movements in Hindusthani music. In: Proceedings of the NSA-2006, New Delhi
5. Datta AK (1996) Generation of musical notations from song using state-phase for pitch detection algorithm. J Acoust Soc India XXIV
6. Datta AK, Sengupta R, dey N, Nag D (2008) Automatic classification of 'meend' extracted from the performances in Hindustani vocal music. Proceedings of FRSM-2008
7. Datta AK, Sengupta R, dey N, Nag D (1997) A methodology for automatic extraction of 'Meend' from the performances in Hindustani vocal music. J ITC Sangeet Research Academy
8. Datta AK, Solanki SS, Sengupta R, Chakraborty S, Mahato K, Patranabis A (2017) Signal analysis of Hindustani classical music. Springer Nature, ISBN 978-981-10-3959-1 (eBook)

Computational Music and Speech

Digital Music, Representation, Data, Machine Learning, Similarity Analysis, Metadata Description, Voice Searching, Libraries

Speech Based Emotion Detection Using R

Nida Farheen and Subarna Chatterjee

Abstract As the human–machine interactions are increasing, there is a growing need ways of communication abilities between humans and machines. Emotion plays a vital role in building the human–machine interactions more original, and it helps in identifying the mood and state of the person. Human emotions impact the perception and everyday activities like communication, learning, and decision making. Speech is the most natural way of communication among human beings across the globe. It is wrapped with many intended meanings and depends on social circumstance and cultural surroundings. As a result, the voice-based call centers (VbCC) are showing rapid growth. Speech processing for emotion detection and recognition is the study of linguistic patterns pertaining to different languages and pronunciations, variations in pitch, tone, and intensity as well the voice quality, also variations in frequency and amplitude of the speech signals. Depending on these parameters, the speech signals are classified into appropriate classes of emotions through various algorithms of pattern recognition and machine learning. An algorithm for emotion detection is proposed here and is implemented in R tool. The algorithm is tested on open speech and language resources (Open SLR) four emotion datasets. Here, three classifiers are used and compared.

Keywords Emotion detection · Speech · R tool · Machine learning

1 Introduction

The human–computer interaction (HCI) is still a challenging issue for the computer systems to interpret and recognize speech with emotions. Human speech contains

N. Farheen
MLIS, Ramaiah University of Applied Sciences, Bangalore, India

S. Chatterjee (✉)
Department of CSE, Ramaiah University of Applied Sciences, Bangalore, India
e-mail: subarna.cs.et@msruas.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_8

linguistic (qualitative features) and paralinguistic (quantitative features) information that can be extracted by speech processing.

In speech processing, the audio wave is first converted into a digital sequence of representation, second pre-processed, and then features are extracted, and after that the audio signal is classified based on the features. Speaker's emotion can be recognized using text-mining approach on audio file after converting it into text. Danisman and Alpkocak [1] have shown that it looks easy for a human to understand and identify the emotion of an audio file, but average score for automatic detection (by algorithms) of different emotion is between 56 and 85%.

It finds applications in human interaction and assistance, digital speech coding, spoken language dialog systems, text-to-speech analysis and automatic speech recognition (ASR) for speaker or gender recognition, information extraction, language or speaker identification or verification, call center automation of voice calls, communication, music compositions, and more. Some real-world examples would be-Google voice assistant, SIRI, Amazon's ALEXA, Cloud vision API by Google, Amazon recognition, Clarifai API, and more.

Shen. et al. [2] have recognized different emotion states using the characteristics: pitch, energy, LPCC, MFCC, and LPCMCC. They have analyzed and differentiate different fusion of features. They have used Berlin emotional database for the experiment.

Speech emotion recognition has potential parametric difficulties, and they are not easy to compute and interpret due to several factors of different parameters and various challenges in computing. It is a difficult task for the computer because of the complexities in different languages and pronunciations. The various parameters and challenges faced are acoustic, prosodic, linguistic, signal to noise ratio, voice activity detection, echo effect, lombard effect, spoken languages, channel variability, speaker gender, phonemes and homonyms, pitch, amplitude, intensity, and frequency.

The organization of the paper is as follows: Sect. 1 comprises the introduction. Section 2 presents review of existing algorithm for emotion detection. Section 3 describes the methodology of the proposed algorithm. Section 4 covers the simulation results with discussion. Section 5 presents the concluding remarks.

2 Existing Algorithms for Emotion Detection

There are various techniques that are used for emotion detection algorithm as mentioned in Table 1. The algorithm implemented here is an acoustat function [3]. This function is used for feature extraction. It returns statistics based on short-term Fourier transform (STFT) and time and frequency contours. STFT works better compared to other algorithms because of following reasons:

1. Complexity: When features are extracted using MFCC [1, 4–6, 8–14] or LPCC algorithms, for each audio sample dimension complexities arise. Dimensionality reduction that needs to resize all the feature matrices is not preferable in MFCC or

Table 1 Analysis and comparison of various techniques for emotion detection

Algorithms	Comparison	Analysis
Hidden Markov Model (HMM) [5, 6, 8, 9, 14]	HMM is most extensively used in emotion analysis. A simple technique based on the principles of Markov chain model takes probability distribution into account. It splits entire speech signal into smaller samples having some probability, and these samples represent the hidden states of the model	HMM is easy to compute, and therefore, it is widely used for classifying the emotions in a speech signal. It uses statistical methods like mean, maximum, minimum, and kappa to calculate the probability of the hidden states in the speech sequence.
Convolutional Neural Network (CNN) [15]	The densely connected neural nets are deeper and efficient to train if they contain shorter connections between input and output layers. It connects each layer in a feed-forward fashion, and features of all previous layers are used as inputs into all subsequent layers	Very effective and accurate algorithm. CNNs aggregate the probabilities of the word sequence by using statistical methods such as mean, median, RMSE, and MAPE to classify the emotions. The accuracy of CNN algorithm is hard to beat as it does both feature extraction and classification, in addition, also predicts based on learning. Hence, it is a very efficient and robust technique.
k-Nearest Neighbors (kNN) [5–7, 14]	It performs classification of the input speech signal by the voting scheme of the nearest vectors available. It works on the posterior probabilities of the class labels	Accuracy calculated by statistics distance metrics like the Euclidian distance measure. Hence, the efficiency is higher.
Mel Frequency Cepstral Coefficient (MFCC) [1, 2, 4–6, 8–14]	Important feature extraction algorithm is used widely for speech recognition. It is a better feature algorithm than linear prediction coding (LPC)	Its accuracy is very high in most speech recognition implementations for emotion recognition.
Linear Prediction Cepstrum Coefficients (LPCC) [2, 6]	Here coefficients of the parameters are considered, and hence, it gives the best results	It gives the best results. The error rate is very minimal, and the error signal is filtered out, hence the accuracy increases.
Expectation Maximization [10]	It is a unique technique that finds the optimal values of the parameters chosen. It is flexible and can be used on GMM, HMM set of observations. It is based on the maximum likelihood principle and is an iterative process, having two steps: E-step and M-step	Provides optimal results. Its accuracy never hinders since it works on the mathematical and statistical principles; hence, it is a robust, effective, and efficient algorithm.
Gaussian Mixture Model (GMM) [6, 10]	Probability density function represented as the weighted sum of Gaussian components densities. GMM is calculated for every emotion. The posterior probabilities are calculated and are compared with the emotion output	It is calculated by the statistical methods of mean and covariance matrix of the PDFs. It depends on the weighted sum of all the GMMs in a model; hence, to always adjust the weights, covariance matrix and mean vector are a tedious task and therefore makes the model less efficient and leads to slow computing.
Vector Quantizer (VQ) [4]	The technique is to map the feature vectors from large feature vector space to a finite region of feature vector space and works on principle of block coding where each region is called as a cluster and can be represented as a code-word	Used for speech coding, speaker recognition, and emotional speech detection as it makes the forms of speech recognition easier and feasible.

LPCC algorithms as it causes curse of dimensionality problem which increases the complexity of the algorithm.

2. Accuracy: if MFCC and LPCC algorithms are implemented and the dimensions are attempted to be resized of all the feature matrices, it will still cause a problem and will affect the feature vector space and therefore reduces the accuracy of the model.

3. Computational speed: MFCC or LPCC algorithms have complex structures; hence, the execution takes a lot of time and thereby decreases the efficiency of the model. On the other hand, the STFT algorithm is simple, has less complex structure, and hence faster at computing.

3 Methodology

3.1 DataSet

The real-time audio samples are taken from an open web source called OpenSLR (openslr.org). Four datasets are downloaded. Each of the dataset represents one emotion- dataset 1—Angry, dataset 2—Happy, dataset 3—Neutral, and dataset 4—Sad. Each of the four datasets has 50 samples each of audio waves of emotions spoken in Javanese, a Native Indonesian language.

The model developed has three parts, firstly to pre-process the audio samples, secondly to train the samples, and thirdly to test the samples for accurate predictions.

3.2 Proposed Algorithm of the Emotion Detection Model

STEP 1: Read audio files—
The audio samples are read from the audio file location using command "list.files" to access the files directly from its location, and "readWave" to read each sample from the audio file. This is done for all the four datasets.

STEP 2: Data preprocessing —
Each audio sample is passed through a windowing technique called Hamming window. Windowing is applied to improve the spectral characteristics of the analyzed signal. We applied the window length equal to 512 samples.

STEP 3: Feature extraction —
The acoustat function is used to extract unique features of the audio wave. It works on the principle of STFT. It returns the time and frequency statistics. It converts the audio wave into frequency and time domain matrix. Resulting matrix will have a time and a frequency curve.

From each curve, the median (M), initial percentile (P1), terminal percentile (P2), and interpercentile range (IPR) are measured. The function returns a list with ten items:
timecurve: This curve is a two-dimensional matrix(x, y), the first column is time(s), and the second column is the amplitude.
freqcurve: This curve is a two-dimensional matrix, first column is frequency (kHz), and the second column is amplitude.
timeP1(initial percentile), **timeM**(median), **timeP2**(terminal percentile), **timeIPR** (interpercentile range) are calculated with respect to time.
freqP1(initial percentile), **freqM**(median), **freqP2**(terminal percentile), **freqIPR** (interpercentile range) are calculated with respect to frequency.
The following eight features are extracted from each sample of all the four audio wave datasets:- **timeP1**, **timeM**, **timeP2**, **timeIPR** and **freqP1**, **freqM**, **freqP2**, **freqIPR**.
A percentile is a measure that gives us what percent of the total frequency of a dataset was at or below that measure. The n-th percentile [17] of a dataset is the value that cuts off the first n percent of the data values when all of the values are sorted from least to greatest. The 90th percentile of a dataset is the value that cuts of the bottom 90% of the data values from the top 10% of data values. The 50th percentile represents the median value of a datasets. Similarly, we can say that the first percentile or initial percentile of a dataset gives the minimum value; last or terminal percentile or 100th percentile of a dataset gives the maximum value.

STEP 4: Data preparation —
After all, eight important features are extracted from each of the dataset, and the result we get from 50 samples of four datasets, having eight features each, gives a total of 50 * 4 = 200 = total of 200 samples.
200 rows X 8 columns, therefore, are the total dimension of all the datasets are 200 * 8, i.e., 200 audio samples and eight extracted features. All the datasets are added row wise by using command "rbind."
The independent output variable is added as the last column, and the classes are assigned labels. Therefore, including the output variable, the total dimension of the whole dataset becomes 200 * 9.
Dataset 1—Angry is being assigned class label as **class 1**
Dataset 2—Happy is being assigned class label as **class 2**
Dataset 3—Neutral is being assigned class label as **class 3**
Dataset 4—Sad is being assigned class label as **class 4**
These class labels define the classes to which the dataset belong when the model is trained and classified. The model when trained and tested for classification, classifies the dataset into respective classes of emotions.

STEP 5: Classification—
Three classification techniques are used in this algorithm. They are classification and regression trees (CART), kNN [5, 7], and random forest (RF). The whole dataset is

partitioned into two with ratio 7:3. That means out of the whole dataset, 70 samples are taken to train the model, and rest 30 are taken to test the model accuracy.

(a) Training: Using 70% of the data, i.e., out of 200 samples, 140 samples are used for training the classifiers. Their performance accuracy is checked by statistical summary.

(b) Testing: If the accuracy is good, then the remaining 30% of the data is tested of the model. Otherwise, there must be some modeling issue with the classifier, and hence, the entire model needs to be changed.

4 Analysis and Results

The audio samples are collected from the Open SLR site which contains four sets of emotion audio samples—Angry, Happy, Neutral, and Sad. Total of 200 samples are considered for the analysis. Each audio sample is read into R-tool, and the samples are restricted to the window length 512 using Hamming window. The results are stored in the matrix of time and frequency curve (Figs. 1, 2, 3 and 4).

Eight attributes are measured from both time and frequency curve. The total size of matrix is 200 * 8. Each audio sample dataset is assigned with a class label 1, 2, 3, and 4, and an output independent variable is added to the matrix, so the total size of matrix changes to 200 * 9.

Figures 1 to 4 depict the time and frequency curve periodogram of each signal sample

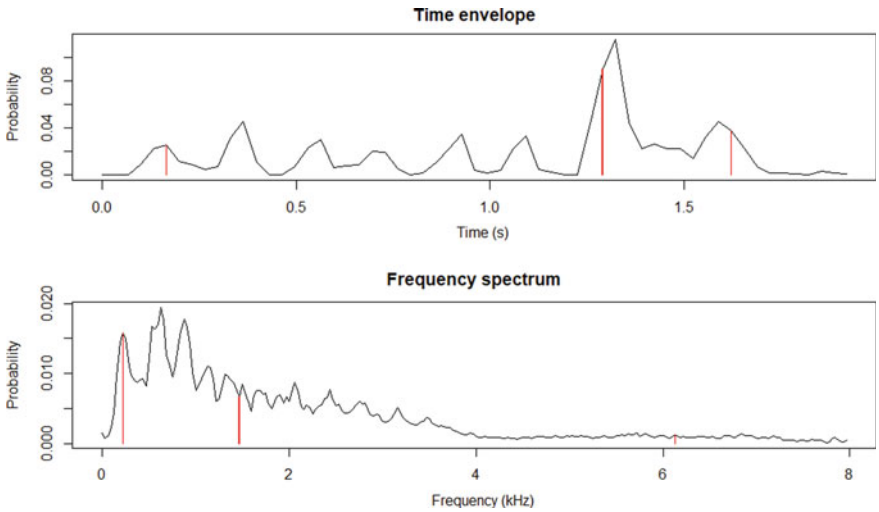


Fig. 1 Time and frequency curve periodogram of each sample from dataset 1—Angry

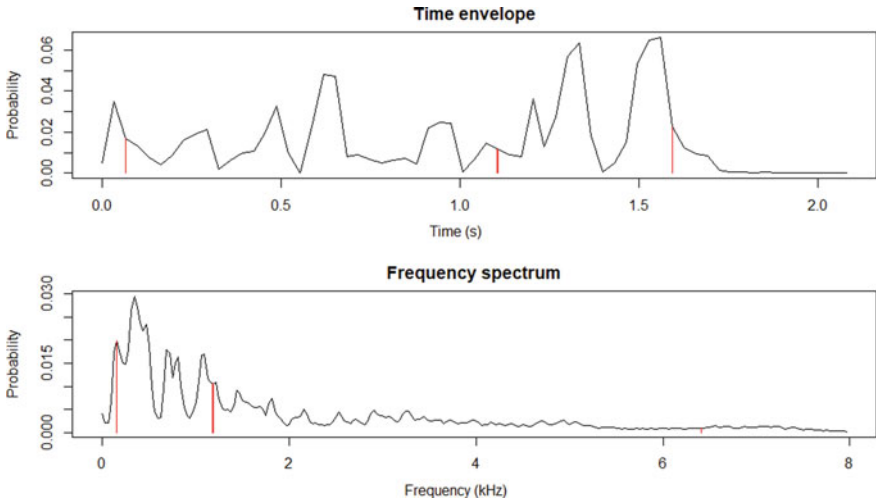


Fig. 2 Time and frequency curve periodogram of each sample from dataset 2—Happy

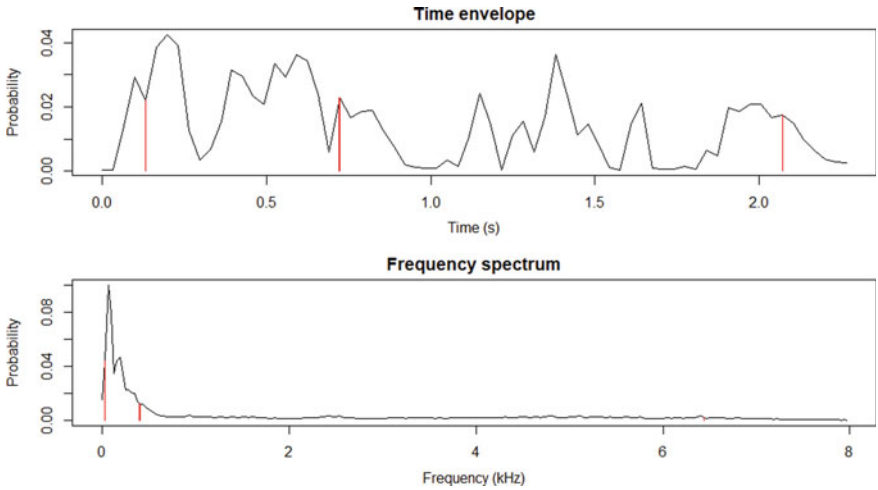


Fig. 3 Time and frequency curve periodogram of each sample from dataset 3—Neutral

of dataset 1-Angry, dataset 2-Happy, dataset 3-Neutral, and dataset 4-Sad using R tool.

The whole dataset matrix (200 * 9) is divided into test and train dataset with the ratio 70:30. The classification is done by the classifiers-CART, kNN, and RF. The performance of the models is evaluated by confusion matrix and statistical analysis.

The analysis for samples from each dataset of emotions is done by statistical summary analysis and confusion matrix. Table 2 shows the statistical summary of the four samples from the four datasets: dataset-1 for Angry, dataset-2 for Happy,

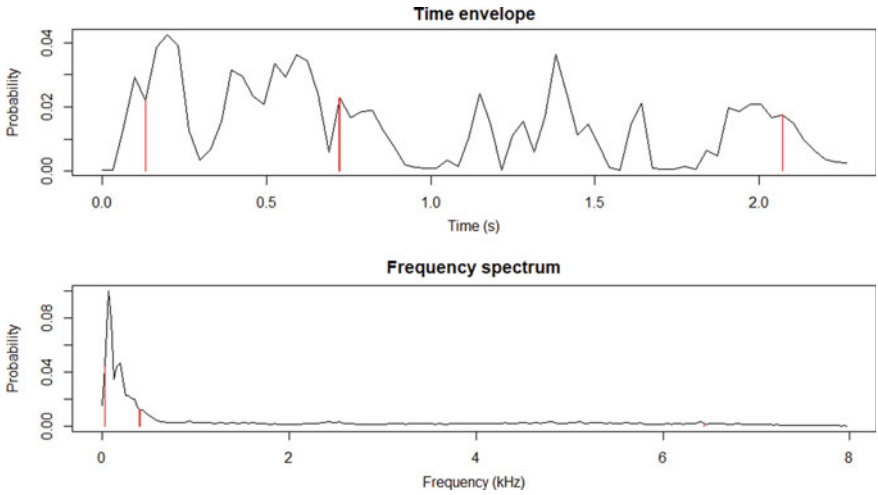


Fig. 4 Time and frequency curve periodogram of each sample from dataset 4—Sad

Table 2 Statistical Summary of the Samples

	No. of Samples	Duration (sec)	Sampling rate(Hz)	Channels (Mono)	PCM (int)	Bits
Wave Object 1	187676	3.91	48000	Mono	TRUE	16
Wave Object 2	172885	3.6	48000	Mono	TRUE	16
Wave Object 3	189000	3.94	48000	Mono	TRUE	16
Wave Object 4	166800	3.48	48000	Mono	TRUE	16
Summary statistics for channel(s)						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Wave Object 1	-24978.000	-582.000	26.000	0.227	1113.000	25822.000
Wave Object 2	-26387.000	-469.000	26.000	1.523	997.000	17834.000
Wave Object 3	-26404.000	-220.000	4.000	1.437	540.000	19842.000
Wave Object 4	-29101.000	-401.000	10.000	2.319	1065.000	16622.000

Table 3 Confusion matrix, Statistical summary, Testing Results of Classifier CART from R tool

Confusion Matrix and Statistics				
	Reference			
Prediction	1	2	3	4
1	15	0	0	0
2	0	15	0	0
3	0	0	15	0
4	0	0	0	15
Overall Statistics				
Accuracy :			1	
95% CI :			(0.9404, 1)	
No Information Rate :			0.25	
P-Value [Acc > NIR] :			< 2.2e-16	
Kappa :			1	
Mcnemar's Test P-Value			NA	
Statistics by Class				
	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	1.00	1.00	1.00	1.00
Specificity	1.00	1.00	1.00	1.00
Pos Pred Value	1.00	1.00	1.00	1.00
Neg Pred Value	1.00	1.00	1.00	1.00
Prevalence	0.25	0.25	0.25	0.25
Detection Rate	0.25	0.25	0.25	0.25
Detection Prevalence	0.25	0.25	0.25	0.25
Balanced Accuracy	1.00	1.00	1.00	1.00

dataset-3 for Neutral, and dataset-4 for Sad. All the wave objects have the same sampling rate 48,000 Hz, all channels are Mono, and all have the same bit value, which is 16. The number of samples present in each wave files and the duration of the waves are different.

Statistical summary of the channel of the same four samples from the four datasets shows different values for minimum, first quartile, median, mean, third quartile, and maximum, where median gives the center of the data; minimum and maximum—the range of the data and first and third quartile gives the sense of spread of data.

Generally, the 25th percentile is called as the first quartile ($Q1$), 50th percentile is the median, and the 75th percentile is called the third quartile ($Q3$).

The classifier model is tested with audio samples from the four datasets (1 to 4), and performance of each model is validated by the accuracy, statistical measurements such as mean, median, and kappa values. Table 3 shows the confusion matrix, statistical summary, and testing results of classifier CART, which are measured using

Table 4 Confusion matrix, Statistical summary, Testing Results of Classifier kNN from R tool

Confusion Matrix and Statistics				
	Reference			
Prediction	1	2	3	4
1	15	0	0	0
2	0	13	0	0
3	0	0	13	0
4	0	0	0	14
Overall Statistics				
Accuracy :			0.9167	
95% CI :			(0.8161, 0.9724)	
No Information Rate:			0.25	
P-Value [Acc > NIR] :			<2.2e-16	
Kappa :			0.8889	
Mcnemar's Test P-Value			NA	
Statistics by Class				
	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	1.0000	0.8667	0.8667	0.9333
Specificity	0.9556	0.9778	0.9778	0.9778
Pos Pred Value	0.8824	0.9286	0.9286	0.9333
Neg Pred Value	1.0000	0.9565	0.9565	0.9778
Prevalence	0.2500	0.2500	0.2500	0.2500
Detection Rate	0.2381	0.2222	0.2222	0.2381
Detection Prevalence	0.2833	0.2333	0.2333	0.2500
Balanced Accuracy	0.9778	0.9222	0.9222	0.9556

R tool. The confusion matrix, statistical summary, and testing results of the three machine learning classifiers CART, kNN, and RF are shown in Tables 3, 4 and 5.

Table 6 shows the comparison of the classifiers CART, kNN, and RF. Two measurements accuracy and kappa values are calculated here for comparison. Table 6 displays the best results for random forest classifier. The accuracy is obtained from the confusion matrix for all the three classifiers. The kappa coefficient (K) [16] is a statistical value which measures inter-rater reliability for qualitative items. Intra-rater reliability is related to the different calculations done by the same person. High value of both accuracy and kappa shows better measurements. The average accuracy shown in Table 6 for the algorithms kNN is 88.84%, CART is 98.48%, and RF is 100%. The average kappa shown in Table 6 for the algorithms kNN is 80.97%, CART is 97.96%, and RF is 100%.

Table 5 Confusion matrix, Statistical summary, Testing Results of Classifier Random Forest from R tool

Confusion Matrix and Statistics				
	Reference			
Prediction	1	2	3	4
1	15	0	0	0
2	0	15	0	0
3	0	0	15	0
4	0	0	0	15
Overall Statistics				
Accuracy :		1		
95% CI :		(0.9404,1)		
No Information Rate:		0.25		
P-Value [Acc > NIR] :		<2.2e-16		
Kappa :		1		
Mcnemar's Test P-Value		NA		
Statistics by Class				
	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	1.00	1.00	1.00	1.00
Specificity	1.00	1.00	1.00	1.00
Pos Pred Value	1.00	1.00	1.00	1.00
Neg Pred Value	1.00	1.00	1.00	1.00
Prevalence	0.25	0.25	0.25	0.25
Detection Rate	0.25	0.25	0.25	0.25
Detection Prevalence	0.25	0.25	0.25	0.25
Balanced Accuracy	1.00	1.00	1.00	1.00

5 Conclusion

The algorithms are implemented using acoustat function and STFT algorithm for emotion recognition. Classifiers used are kNN, CART, and RF. The validation and analysis for each of the classifier are compared by measuring the accuracy, statistic measure such as mean, median, and kappa values. The average accuracy of the algorithms kNN is 88.84%, CART is 98.48%, and RF is 100%. The average kappa of the algorithms kNN is 80.97%, CART is 97.96%, and RF is 100%.

Future scope of this work can be:

- Using more robust algorithms such as SVM, HMM, deep learning algorithms to train the model (deep neural networks, recurrent neural network).
- Training the classifier with larger datasets to check complexity.
- Taking samples from different languages to get high variability in the dataset.

Table 6 Statistical summary of Accuracy results of classifiers CART, kNN and Random Forest

Models: kNN, CART, Random Forest							
Number of resamples: 30							
Accuracy							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Average
kNN	0.6923077	0.8595238	0.9230769	0.8886355	0.9285714	1	85.84%
CART	0.9285714	1.0000000	1.0000000	0.9953968	1.0000000	1	98.48%
Random Forest	1.0000	1.0000000	1.0000	1.0000000	1	1	100%
Kappa							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Average
KNN	0.5873016	0.8092565	0.8968254	0.8506104	0.9045988	1	80.97%
CART	0.9041096	1.0000000	1.000000	0.9938275	1.0000000	1	97.96%
Random Forest	1.0000000	1.0000000	1.000000	1.0000000	1	1	100%

References

1. Danisman T, Alpkocak A (2008) Emotion classification of audio signals using ensemble of support vector machines. Percept Multimodal Dialogue Syst 5078:205–216
2. Shen P, Changjun Z, Chen X (2011) Automatic speech emotion recognition using support vector machine. In: International conference on Electronic & Mechanical Engineering and Information Technology, pp 621–625
3. Fristrup KM, Watkins WA (1992) Characterizing acoustic features of marine animal sounds. Woods Hole Oceanographic Institution Technical Report WHOI-92-04
4. Swain M, Routray A, Kabisatpathy P (2018) Databases, features and classifiers for speech emotion recognition: a review. Int J Speech Technol 21(1):93–120
5. Ayadi ME, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44:572–587
6. Krothapalli SR, Koolagudi SG (2013) Speech emotion recognition: a review. In: Emotion recognition using speech features. Springer, New York, pp 15–34
7. Nannapaneni R, Chatterjee S (2020) Human emotion recognition through facial expressions. In: MISS2k20, Springer Algorithms for Intelligent Systems (AIS) Series, ISSN: 2524-7565
8. Rabaoui A, Davy M, Rossignol S, Lachiri Z, Ellouze N (2007) Using one-class SVMs and wavelets for audio surveillance systems. IEEE Trans Inf Forensic Security
9. Morton KD (2010) Bayesian techniques for adaptive acoustic surveillance. Ph.D. Thesis, Duke University, (2010)
10. Huang C, Song B, Zhao L (2016) Emotional speech feature normalization and recognition based on speaker-sensitive feature clustering. Int J Speech Technol 19:805–816
11. Godsill SJ, Rayner PJW (1998) Digital audio restoration—a statistical model based approach
12. Zwiecki MN, Ciołek M, Cisowski K (2015) Elimination of impulsive disturbances from stereo audio recordings using vector autoregressive modeling and variable-order Kalman filtering. IEEE Trans Audio Speech Lang Process 23(6):970–981
13. Crocco M, Cristani M, Trucco A, Murino V (2016) Audio surveillance: a systematic review. ACM Comput Surv 48(4), 52:1–52:46
14. Arslan Y, Canbolat H (2017) A sound database development for environmental sound recognition. Signal Processing and Communications Applications Conference (SIU), 25th, 2017

15. AyadHuang G, Liu Z, Maaten LVD, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, pp 2261–2269
16. Zaiontz C. Cohen's Kappa. http://www.real-statistics.com/reliability/interrater-reliability/cohens-kappa/
17. https://www.statology.org/percentiles-in-r/

Vocalist Identification in Audio Songs Using Convolutional Neural Network

Manoj Kaushik, Shikha Rani, and Vrinda Yadav

Abstract There are millions of songs in the world. Manually, it is a cumbersome task to classify songs according to the vocalist. On the other hand, search algorithms can efficiently recognize a song but these algorithms use pattern matching technique in a database and therefore cannot be generalized over known songs. Music companies often face these types of challenges in automation of song classification. Hence, there is a need for artificial intelligence-enabled method to classify songs based on vocalists which can also be generalized for unknown songs. This paper presents a deep learning architecture to achieve the objective of identifying vocalists in audio songs. For that purpose, a novel bollywood song dataset (BSdataset) is created. The pre-processing of dataset involves the conversion of an audio file into a spectrogram, i.e. visual representation of frequencies of audio signal as it varies with time and then uses these spectrograms as an image to train a convolutional neural network (CNN) for classification of vocalists in an audio song.

Keywords Vocalist classification · Music · Songs · Deep learning · Convolutional neural network

1 Introduction

Conventionally, music retrieval can be achieved through fix pattern matching against music databases. It requires search algorithms [13] and hence cannot be generalized to identify a song like human beings. We can identify songs by listening to them and can even identify the vocalist in it.

In earlier approaches, machine learning methods have employed using low-dimensional feature representation for vocalist classification in audio signals [1, 2]. Summarization of audio signal is needed because conventional models bear the curse of dimensionality; hence, dimensionality reduction is essential [3], so that high-

M. Kaushik (✉) · S. Rani · V. Yadav
Department of Computer Science and Engineering,
Centre for Advanced Studies, Lucknow 226031, India
e-mail: vrinda@cas.res.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_9

dimensional signals can be processed with less computing power. Song classification based on vocalist is a challenging task because music is mingled with the vocals and the voice of the vocalist changes with time.

The major difference between the musical information retrieval (MIR) and deep learning-based approach is that correct representation of audio data is difficult to detect in MIR. The previous research was focused on vector summaries of frequencies in an audio clip, precisely mel-frequency cepstral coefficient (MFCC) because it represents both timbral information and key in an audio file [4]. The trouble with MFCC and other popular features like real-time features, pitch content features, rhythmic content features and whole file [5] is that summarization of a short time duration audio clip into vector results in the loss of temporal structure.

In this study, we used the CNN architecture for audio files (songs) classification as mentioned [6]. In a country like India where bollywood music is very popular but there is almost no research work on identification of singers in Indian bollywood songs to the best of our knowledge. However, some work is done on mood detection in songs [14].

The rest of the paper is categorized as follows. Section 2 discusses the related work. Section 3 discusses the methodology and the architecture of the CNN model. In Sect. 4, results are discussed, which leads to the conclusion and future work in Sect. 5.

2 Related Work

This section discusses related work in the area of audio processing for music retrieval. In particular, the following four methods discussed are as follows: (i) machine learning approaches in music processing, (ii) baseline for artist classification, (iii) audio representation as spectrograms and (iv) convolutional neural networks. These methods are detailed below.

2.1 Machine Learning Approaches in Music Processing

State-of-the-art research work related to sound is more focused on speech recognition, speech translation and event noticing [7]. In the era of cloud-based hosting of music, the machine learning relevance in classification, categorization and recommendation of personalized songs has increased. This technology is also useful in genre classification [8]. Feature extraction (like MFCC) in machine learning is an pre-processing task while it is automated in CNN models [9]. Although deep learning has immense scope in MIR, less research work has been done on vocalist classification unlike genre classification.

2.2 Baseline for Artist Classification

The use of machine learning algorithms and deep learning architectures in vocalist classification is limited [10]. Other previous work is nearly a decade old, and the applied algorithms are not suitable for high-dimensional data. For example, Whiteman et al. [11] have used the MFCC feature of one second audio signal on support vector machine (SVM) classifier with an accuracy of 50% while the dataset is not available in public domain. They have used 240 songs with multiple genres.

The dataset **artist20** [15] is released by LabROSA, applied Gaussian classifier with full covariance to make a bottom line of vocalist classification. MFCC features randomly sampled from every vocalist, and model demonstrates accuracy of 56%. The final accuracy achieved is 59% after including a few handcrafted features in the model [2].

2.3 Audio Representation as Spectrograms

Classical music retrieval tasks are highly dependent on MFCC features in which frequency content is extracted from an audio frame of a small window. Although, recent work involves the use of spectrograms to extract temporal structure and with better results in audio classification [12]. A spectrogram is a visual representation of the audio signal. It is the squared magnitude of short-time Fourier transform (STFT) of an audio signal. The equation of discrete STFT is given in Eq. (1) below:

$$STFT\{x(n)\}(m,\omega) = X(m,\omega) = \sum_{n=-\infty}^{\infty} x(n)w(n-mR)e^{-j\omega n} \tag{1}$$

where
$x(n)$ = input signal at time n
$w(n)$ = window function of length M
$X(m,\omega)$ = DTFT of windowed data centred about time mR
R = size of hop, in samples, between successive DTFTs

Figure 1 shows the one-minute spectrogram of an audio file from BSdataset which includes both frequency content and temporal variation. Spectrograms are difficult to interpret manually but CNN can identify patterns in spectrogram images using various filters. CNN filters can extract features from the image representation of spectrograms.

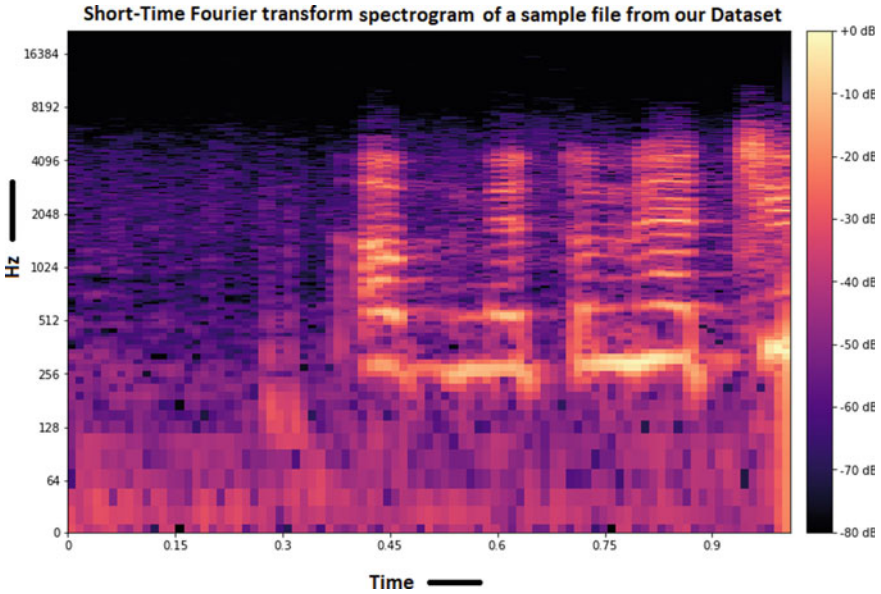


Fig. 1 Spectrogram of a song from BSdataset

2.4 Convolutional Neural Networks

To use CNN model in vocalist classification, we have firstly converted all voice segments into spectrograms and then trained CNN model for vocalist classification. Hershey et al. mentioned the usage of CNNs to perform sound classification using spectrograms [18]. Even though CNN is used for image classification but nowadays, researchers are also using CNN for audio clasiification.

Audio clips are pre-processed before training the CNN model. The pre-processing module converts audio clips into spectrograms using STFT. CNN can extract features from spectrograms instead of raw audio signals. CNN is helpful for MIR-related tasks where CNN layers function as feature extractors.

3 Methodology

The following sections discuss the BSdataset used in the proposed work, pre-processing of audio signals and the architecture of the CNN model.

Table 1 BSdataset specification

Property	Value
Total no. of clips	122 * 2 = 244
Total no. of singers	4
Bitrate	32 kbps
Sample rate	44.1 kHz

3.1 Dataset

The BSdataset is created from 122 audio songs of four Indian bollywood singers. Duration of each song varies from 3.5 min to 5 min. Two, one-minute length audio segments are extracted from each song. Both audio segments are extracted by skipping 40 s from the beginning and the end. So, there are total 244 (122*2) audio clips in the dataset. The specification of BSdataset is shown in Table 1. This dataset contains four classes, and each class belongs to the name of the vocalist. Apart from the BSdataset, a seprate testing set containing 74 one-minute audio clips is also prepared for testing proposed CNN model.

The BSdataset is divided into training and validation sets. The training set is 75%, and the validation set is 25% of the complete dataset. Training and validation sets contain 183 and 61 audio files, respectively. Labelling is done through OneHotEncoding [16], which converts the categorical labels into numerical labels using array of 0s and 1s. OneHotEncoding is more expressive for CNN than categorical labels.

3.2 Pre-processing of Audio Signals

For each audio file, the sampling rate (sr) 44,100 Hz, length of fast Fourier transform (FFT) 2048 Hz and the number of samples between successive frames are 512 and then trimmed the leading and trailing silence from an audio signal with a threshold of 60 decibels. Normalization is done on trimmed data followed by data padding with trailing zeros. STFT is applied to padded data. STFT is a representation of a signal in the time-frequency domain, and the time-frequency representation is computed by discrete Fourier transforms (DFT) over short overlapping windows. The obtained STFT data which is an amplitude spectrogram is converted to a db-scaled spectrogram with float data type instead of integer data type to obtain more precision.

The next step involves data transformation to replace the missing values (NaN) with the mean values of columns. The transformed data is a db-scaled matrix, and its dimensions have expanded along axis-0 and axis-3. Figure 2 shows the pre-processing flow of audio signals.

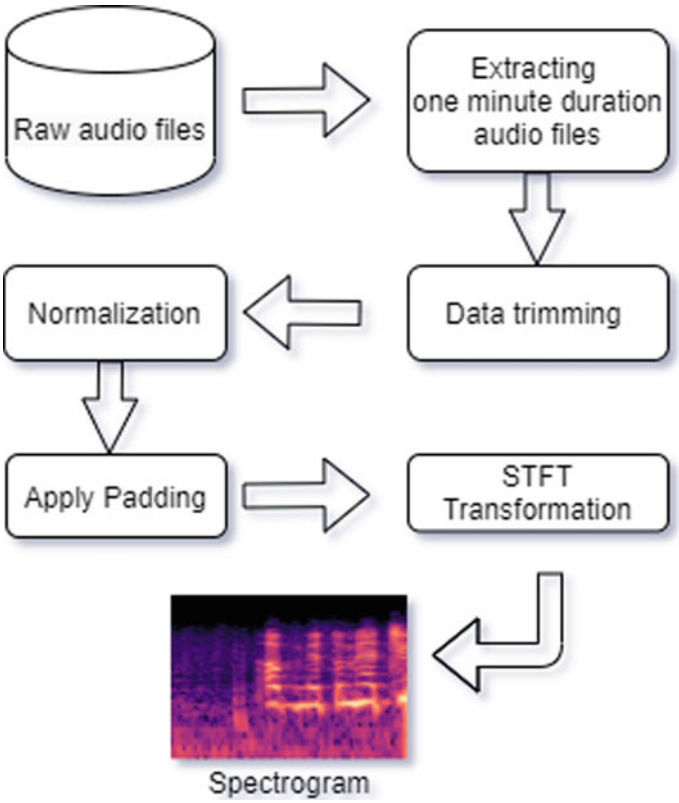


Fig. 2 Pre-processing of dataset

3.3 Architecture of the CNN Model

One of the objectives of proposed work is to substitute the usual MFCC feature-based approach with the spectrogram. The model contains two convolutional layers, two max pooling layers and two dropout layers so that computation complexity can be reduced, one flatten layer which flattens all the 81,280 neurons in a single layer. The last two layers of all 81,280 neurons are densely connected to the 16 neurons of the next dense layer. These 16 neurons are further connected to the last output layer with only 4 neurons as representing 4 classes.

There are 16 convolutional filters in each convolutional layer with (1025, 87, 1) input shape, hyperparameters values for padding and stride are set as 'valid' and 1, respectively, and hence, the output shape of first Conv2D layer is (none, 1023, 85, 16) where None represents the batch size. The pool size in the max pool layer is 2×2, and the dropout [19] rate is 40% in the first dropout layer where 45% in the second dropout layer so that number of parameters can be reduced. Figure 3 shows the various layers of CNN architecture. Sigmoid activation function is used

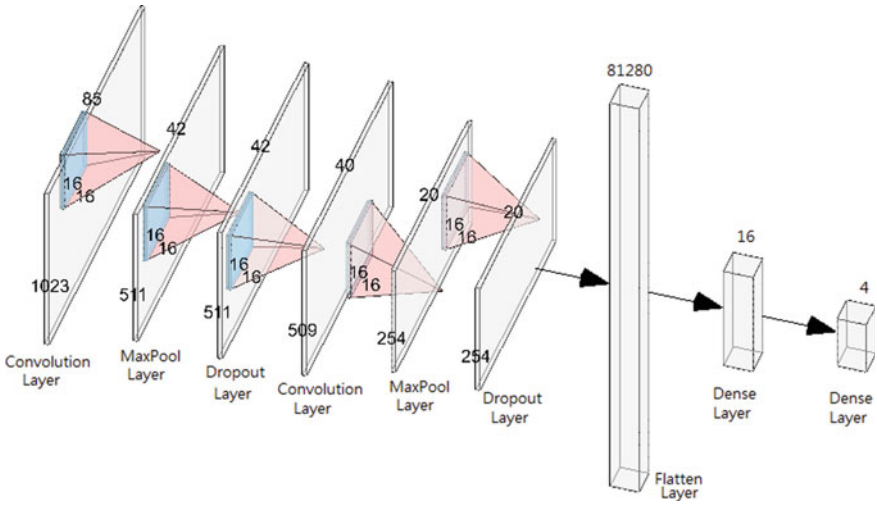


Fig. 3 Visualization CNN architecture

Table 2 CNN Model architecture

Type of layer	Output_Shape	No. of parameters
Conv2D_Layer-1	(None, 1023, 85, 16)	160
MaxPool2D_Layer-1	(None, 511, 42, 16)	0
Dropout_Layer-1	(None, 511, 42, 16)	0
Conv2D_Layer-2	(None, 509, 40, 16)	2320
MaxPool2D_Layer-2	(None, 254, 20, 16)	0
Dropout_Layer-2	(None, 254, 20, 16)	0
Flatten_Layer	(None, 81280)	0
Dense_Layer-1	(None,16)	1300496
Dense_Layer-2	(None, 4)	68

in the first dense layer while softmax activation function is used in the output layer to classify an audio file. The model architecture with the number of parameters is shown in Table 2.

It consists of a total of 1,303,044 trainable parameters. The loss function we have used is categorical crossentropy which is given by Eq. (2).

$$L(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij} * log(\hat{y}_{ij})) \tag{2}$$

where y represents the true value, and $\hat{y}$ represents predicted value.

Table 3 Validation Accuracies in BSDataset

Classes	Training set	Validation set	Validation set accuracy (%)
Kishore	51	17	52.94
Lata	50	16	68.75
Mukesh	39	14	42.85
Rafi	43	14	64.28
Total	183	61	57.205

For the optimization, we have used stochastic gradient descent (SGD) [17]. SGD is an iterative function that is used to optimize with suitable smoothness properties.

4 Results and Discussion

The CNN model is trained on converted audio clips of 1 min duration length into spectrogram. The model is trained over 30 epochs, 350 batches (steps per epoch) for training generators and 150 batches (validation steps) for validation generators. The model training time is 2 h 51 min and 23 s.

Model is trained on the device having two Intel Xeon E5-2650v3 @ 2.30 GHz with 20 cores Processor, 32 GB RAM, GPU of NVIDIA Tesla K40m with 24 GB memory; development environment is Google Colab, and programming language is Python with Tensorflow and Keras deep learning libraries [20].

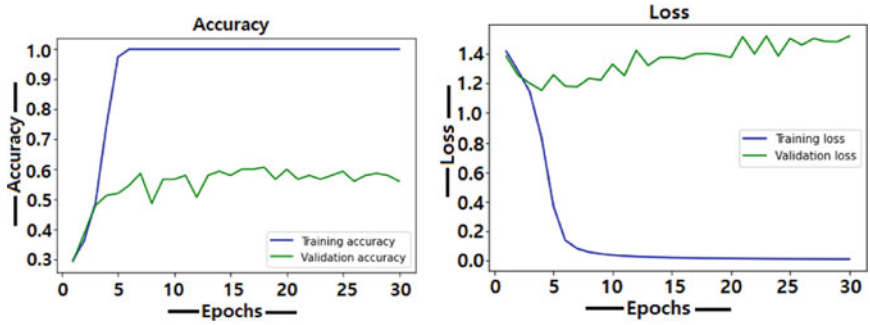
The BSdataset description given in Table 3 is showing the all four classes, number of training and validation samples taken and obtained validation accuracies of each class.

The CNN model achieved overall 60.656% accuracy for test set and 57.205% validation accuracy for BSdataset as shown in Table 3 for vocalist classification. The loss and accuracy are shown in Fig. 4 for 30 epochs training. From Fig. 4a, b, we can see the loss and the accuracy of the model while training over the time.

The confusion matrix shown in Fig. 5 shows how many correct and incorrect predictions are made by the model in validation for all the classes. All predictions for 4 classes can be seen in Fig. 5.

5 Conclusions and Future Work

This paper focuses on vocalist identification in Indian bollywood songs. In this paper, we have employed a deep learning convolutional neural network architecture to identify vocalists from four vocalist categories. Also, we attempted to overcome the



(a) Accuracy (b) Loss

Fig. 4 Training over 30 epochs

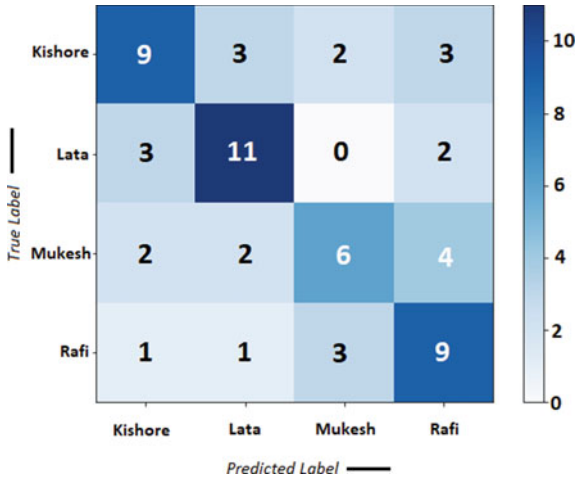


Fig. 5 Confusion Matrix of four classes

high dimensionality problem caused by MFCC features when applied with various machine learning techniques. A deep neural network is capable of handling such problems because feature extraction is convenient in convolutional neural networks. The proposed method can be used in music recommender systems to recommend songs based on singers.

Future work includes enhancement in the convolution neural network layers so that the accuracy of the model can be improved further. The achieved testing accuracy of the model is 60.656% while the validation accuracy is 57.205%. Proposed work can be further extended by incorporating more audio clips of different Indian bollywood singers in the dataset. The audio duration considered in the paper is one-minute length, which can impact performance of the CNN model.

References

1. Whitman B, Flake G, Lawrence S (2001) Artist detection in music with minnow match. In: IEEE signal processing society workshop. IEEE, pp 559–568
2. Ellis DP (2007) Classifying music audio with timbral and chroma features. In: International Society for Music Information Retrieval Conference (ISMIR), vol 7, pp 339–340
3. Verleysen M, François D (2005) The curse of dimensionality in data mining and time series prediction. In: International work-conference on artificial neural networks. Springer, Berlin, pp 758–770
4. Li TL, Chan AB (2011) Genre classification and the invariance of mfcc features to key and tempo. In: International conference on MultiMedia modeling. Springer, Berlin, pp 317–327
5. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Process 10(5):293–302
6. Choi K, Fazekas G, Sandler M, Cho K (2017) Convolutional recurrent neural networks for music classification. In: International conference on acoustics, speech, and signal processing. IEEE, pp 2392–2396
7. Deng L, Yu D et al (2014) Deep learning: methods and applications. Foundations and trends® in signal processing, pp 197–387
8. Choi K, Fazekas G, Sandler M (2016) Explaining deep convolutional neural networks on music classification. arXiv preprint arXiv:1607.02444
9. Lee H, Pham P, Largman Y, Ng AY (2009) Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in neural information processing systems, pp 1096–1104
10. Nasrullah Z, Zhao Y (2019) Music artist classification with convolutional recurrent neural networks
11. Whitman B, Flake G, Lawrence S (2001) Artist detection in music with minnowmatch. In: IEEE signal processing society workshop. IEEE, pp 559–568
12. Choi K, Fazekas G, Sandler M (2016) Automatic tagging using deep convolutional neural networks. arXiv preprint arXiv:1606.00298
13. Wang A et al (2003) An industrial strength audio search algorithm. In: International Society for Music Information Retrieval Conference (ISMIR), vol 2003. Washington, DC, pp 7–13
14. Patra B, Das D, Bandyopadhyay S (2013) Automatic music mood classification of Hindi songs
15. Music Artist Identification: artist20. https://labrosa.ee.columbia.edu/projects/artistid/. Last Accessed 9 May 2020
16. Sklearn preprocessing OneHotEncoder. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html. Last Accessed 9 May 2020
17. Ruder S (2016) An overview of gradient descent optimization algorithms. ArXiv Preprint ArXiv:1609.04747
18. Hershey S, Chaudhuri S et al. CNN architectures for large-scale audio classification. arXiv:1609.09430v2
19. Srivastava N et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res
20. Chollet F et al (2015) Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Swaragram: Shruti-Based Chromagram for Indian Classical Music

Yeshwant Singh, Ranjeet Kumar, and Anupam Biswas

Abstract Computational musicology has seen many advances in the recent past, and researchers have developed different types of tools for analyzing sound, especially with music at the center of attention. These techniques are aimed at western music and somewhat fail to capture the aesthetic of Indian classical music. In this article, we present Swaragram, a feature extraction technique inspired by the Chromagram for the spectral analysis of Indian classical music. We have released our Python implementation of the Swaragram under the GNU-GPLv3 license [1] to help learners of Indian music and encouraging other researchers to use it, change it as they need. Finally, we have applied this tool showing its usefulness in predicting raga in Indian classical music.

Keywords Indian classical music · Shruti · Spectral analysis · Chromagram

1 Introduction

The scientific community has achieved many advances for the analysis of Western music [2], and researchers have documented them very well [3–6]. However, Indian music remains unexplored, and very little computational analysis [7] has been done to completely understand its intricacies. Classical music in the Indian sub-continent has a very long history, from the description in the Vedas [8] to the latest trends of Bollywood music. The theory and understanding of classical music in India have always followed the Guru–Shishya (teacher–student) pedagogy [8], where the Guru (teacher) would pass on his teachings to his/her Shishya (student) through vocal or

Y. Singh (✉) · R. Kumar · A. Biswas
Department of Computer Science and Engineering, National Institute of Technology, Silchar, Assam 788010, India
e-mail: yeshwant_rs@cse.nits.ac.in

R. Kumar
e-mail: ranjeet_rs@cse.nits.ac.in

A. Biswas
e-mail: anupam@cse.nits.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_10

instrumental training. Western music has score notation, which is a way of music documentation, for educational and recreating purposes. Contrary to this, no written notation has been developed for Indian classical music. Before a performance, Indian musicians convey their ideas to each other about what they are going to perform by using concepts like tala (rhythm) and raga structure (underlying melody) [7].

A Saptak (octave) in Indian classical music is split up into 22 parts, each called Shruti [9]. The first Shruti, called Shadja (S), is the tonic of the music, while other Shrutis get their relative positions depending on the Shadja. This gives lots of freedom to the performers of Indian classical music. The Shadja is generally played by an instrument like the tabla or the tanpura for performers' reference. Table 1 gives us the corresponding ratio between the Shadja and any other Shruti in a Saptak. For example, the ratio between the 14th Shruti, called Pancham (P), and Shadja (the 1st Shruti) is 3/2. So, if we take Shadja to be 100 Hz, then Pancham would be 150 Hz. These 22 Shrutis are clubbed together to get 12 swaras, called Sargam and notated as "$S\ r\ R\ g\ G\ M\ m\ P\ d\ D\ n\ N$" which would be somewhat closer to the Western notation of notes "$C\ C_{\#}\ D\ D_{\#}\ E\ F\ F_{\#}\ G\ G_{\#}\ A\ A_{\#}\ B$". When a musician performs

Table 1 Frequency ratios of 22 Shrutis

No.	Shruti name	Symbol	Ratio
1.	Shadja	S	1/1
2.	Ati Komal Rishabh	r1	256/243
3.	Komal Rishabh	r2	16/15
4.	Shuddha Rishabh	R1	10/9
5.	Teevra Rishabh	R2	9/8
6.	Ati Komal Gandhar	g1	32/27
7.	Komal Gandhar	g2	6/5
8.	Shuddha Gandhar	G1	5/4
9.	Teevra Gandhar	G2	81/64
10.	Shuddha Madhyam	M1	4/3
11.	Ek shruti Madhyam	M2	27/20
12.	Teevra Madhyam	m1	45/32
13.	Teevratam Madhyam	m2	729/512
14.	Pancham	P	3/2
15.	Ati Komal Dhaivat	d1	128/81
16.	Komal Dhaivat	d2	8/5
17.	Shuddha Dhaivat	D1	5/3
18.	Teevra Dhaivat	D2	27/16
19.	Ati Komal Nishad	n1	16/9
20.	Komal Nishad	n2	9/5
21.	Shuddha Nishad	N1	15/8
22.	Teevra Nishad	N2	243/128

classical music, the underlying raga's grammar specifies which Shrutis the musician should perform. These selected Shrutis become the swaras (notes) of that raga. This is opposite to Western music that follows the equal-tempered scale—an octave is divided into 12 semitones and the ratio of the adjacent semitones is constant and is equal to $2^{\frac{1}{12}}$ [10].

A Chromagram is a spectral analysis technique that is good for analyzing the harmonic nature of music, and it is extensively used in the field of music information retrieval for various purposes [11–16]. However, it fails to capture the dynamic nature of swaras (notes) in Indian classical music, as our results suggest. This is because their frequencies change depending on the Shadja of the music. We need a flexible and dynamic-natured Chromagram based on Shrutis to analyze Indian classical music. In this article, we describe Swaragram as a technique inspired by the Chroma feature that relies on the Shadja and Shrutis for its computation. This article is split into five sections. The section after this describes the feature extraction process, followed by a brief description of the tool in Sect. 3. Toward the end, we present an application of this tool in Sect. 4, and at the end, in Sect. 5, we present some ideas to extend this work.

2 Feature Extraction

In this section, we present the building blocks for this approach. Some of the feature extraction steps can be replaced by alternative approaches like STFT with CQT transform. Other smoothing techniques can be added to the end of the feature extraction process to make the representation more robust to temporal and tonal variations.

2.1 Tonic Identification

The basis of Swaragram is the principle of Shadja (tonic) variation in Indian classical music. The performer selects a Shadja (tonic) swara frequency for a performance, and this acts as a reference for him/her throughout the act; it is used for tuning all other instruments. As Shadja is the root of all other swaras, they are sung or played relative to it. The selection of Shadja can be anything within the performer's ability and is different from Western music, where the key is fixed (such as A or E). Tonic identification has been studied in the past [17, 18], and they have satisfactory results to be employed to automate this portion of the toolbox.

2.2 Relative Swara Frequency

After identifying the tonic, we need to find out the center frequencies of swaras in Saptaks (octave). The frequency values of Mandra (lower) S and Taar (higher) S are in ratio 2 with the Madhya (middle) S. Other swaras frequency values are calculated based on Table 1. As the change in tonic would result in a completely different set of frequency values for the MIDI pitch values, they are calculated for a given tonic. These 22 frequency ranges are aggregated into 12 ranges, this is due to fact that five swaras "R G M D N" out of seven "S R G M P D N" have two versions as Komal and Shudha, or Shudha and Teevra in the case of M. Each of these versions has two further variations as shown in Table 1, making it $5 * 2 * 2 + 2 = 22$. The S and P are Achal (immovable) swaras —they have a fixed frequency value.

2.3 Logarithmic Frequency Pooling Based on Shruti

Human perception of music is logarithmic in nature, and we hear pitches which are a Saptak (octave) apart to be similar. For this purpose, we use MIDI pitch values from 0 to 127—taking 12 values in a Saptak—to generate the corresponding frequency values for other swaras relative to the MIDI pitch value in an octave. To get this representation from a Spectrogram, the idea is to assign each spectral coefficient $X(m, k)$ calculated via STFT to a swara group, with the center nearest to the calculated frequency based on Table 1. It is given by following Eq. 1 and can found here in detail [6]:

$$P(p) := \{k : F_{\text{MIDI}}(p - 0.5) <= F_{\text{coef}}(k) < F_{\text{MIDI}}(p + 0.5)\} \tag{1}$$

while the bandwidth of a pitch is given by Eq. 2:

$$BW(p) := F_{\text{MIDI}}(p + 0.5) - F_{\text{MIDI}}(p - 0.5) \tag{2}$$

The bandwidth for the lower pitch values turns out to smaller. Equation 1 gives us the frequency frame from the Spectrogram, which needs to be pooled in pitch p. We apply a bandwidth function BW with the pooling to have better tolerance.

Based on the set $P(p)$ we get from Eq. 1, we obtain a log-frequency Spectrogram for MIDI pitches [0: 127] using a simple pooling procedure:

$$Y_{LF}(m, p) := \sum_{k \in P(p)} |X(m, k)|^2. \tag{3}$$

This method logarithmically subdivides the axis representing frequency, labeled linearly according to MIDI pitches.

2.4 Swara Binning

The final step to get Swaragram out of the log-frequency Spectrogram is to accumulate the bands of corresponding pitches that just vary by one or more octaves. Humans perceive the swaras in a periodic sense, which means that humans perceive two swaras to be similar in their tonal texture (harmonic nature of music) if they only differ by an octave [19].

All the spectral domain data that corresponds to a given swara class is aggregated into one group. Given a log-frequency Spectrogram based on MIDI pitch:

$$Y_{LF} : S = [0 : 127]$$

A Swaragram representation:

$$X = [0 : 11]$$

Both Y_{LF} and X are positive real-valued and can be derived by accumulating all pitch coefficients that belong to the same swara class.

3 Toolbox

The parts of feature extraction as explained in Sect. 2 form the basis of our Python implementation [1] and have been released under the GNU-GPLv3 license. The implementation is divided into two submodules (`core` and `utils`) to keep it manageable and following good practices. The `core` submodule is a collection of necessary functions for extracting Swaragram, while the `utils` submodule provides helper and wrapper functions. Table 2 gives a description of Python functions along with their descriptions. Full details on the default parameters and the API can be obtained here [1]. A user can change default values passed to the function parameters to get desired effects.

To show how our released implementation can be used, a sample code example is also included in the released code base and is shown in Table 3. The code example starts with multiple import statements, followed by loading the audio using `utils.loadAudio` with just a wrapper function for `load` from the `librosa` [20] package. It loads the audio in the mono version sampled 22050 Hz, and these defaults can be changed in the `loadAudio` function. In line 8, we calculate the tonic of audio by calling `utils.getTonic`, which is using the `essentia` [21] library for calculating the tonic. In line 10, we calculate the STFT of the audio signal, followed by the reference MIDI pitch and the corresponding frequency related to the tonic calculated previously. Swaragram is computed in line 14 by passing the STFT, the reference pitch, and the reference frequency. The `computeSwaragram` internally does the following things:

Table 2 Overview of the Python implementation API

Function	Description
`Utils submodule`	
`computeRefPitch`	Computes tonic pitch and its corresponding frequency
`computeSTFT`	Calculates the STFT of an audio signal
`getTonic`	Gets the tonic of the audio signal
`loadAudio`	Loads the audio sample
`logCompression`	Log compression of the input value
`Core submodule`	
`swaraRange`	Computes the lower and upper frequency/ies of a MIDI pitch based on shurti scale
`poolMask`	Computes the set of frequency indices that are assigned to a given pitch
`computeSwaragram`	Computes a Swaragram of a Spectrogram
`logFreqSpec`	Computes a log-frequency Spectrogram
`plotSwaragram`	Plots Swaragram

Table 3 Code example

```
from swaragram import core, utils
import librosa
import numpy as np

# load audio file
x, sr = utils.loadAudio('sample.wav')
# calculate tonic of the indian classical music sample
tonic = utils.getTonic(x)
# calculate STFT with win_length=2048 and hop_length=512
stft = utils.computeSTFT(x)
# compute reference pitch for calculating Swaragram
refPitch, refFreq = utils.computeRefPitch(tonic)
# calculate Swaragram for a given ICM sample
swara = core.computeSwaragram(stft, refPitch=refPitch,
                              refFreq=refFreq)
# Get time values for plotting Swaragram
t = librosa.frames_to_time(np.arange(stft.shape\cite{ch10bharatmuni}),
                              n_fft=2048)
# Log compress Swaragram
swara = utils.logCompression(swara)
# Plot Swaragram
core.plotSwaragram(swara, t)
```

1. Calculating the power spectrum.
2. Computing the log-frequency Spectrogram using `core.logFreqSpec`. The computation of the log-frequency Spectrogram involves these steps:
   (a) Calculating the pitch harmonic indices based on the Shruti scale.
      I. Computing the lower and upper-frequency range of swara based on the Shruti scale using `core.swaraRange`
      II. Calculating a mask of indices which are to be accepted using `core.poolMask`.
   (b) Binning them as humans perceive music (based on the calculated indices).
3. Aggregating indices columns based on swara class value.

Next, in line 16, time vector `t` is calculated using `librosa` for plotting the Swaragram. The result of the `computeSwaragram` is log compressed to make it more tolerant to global variations using the `logCompression` function and then passed on to a plotting function from our implementation package, as shown in line 20.

4 Application for Indian Classical Music

4.1 Raga Classification

An audio analysis of Indian music recordings for the raga [22] identification and to automate this process is a major contribution to music information retrieval in the context of Indian music. Most of the raga identification procedures employ pitch-class distributions, while some use vector space modeling and melodic surfaces [23–25]. Western music has seen extensive uses of harmonic analysis for the musical analysis (especially melodic) of audio recordings. The Chroma feature is used in many applications in Western music. In our observation, the effectiveness of the Chroma feature for Indian classical music is inadequate. We performed a small experiment on the Dunya dataset [26], choosing 16 raga randomly out of 203 available in the dataset. We extracted Chroma and Swaragram as two sets of features from about 20 h of music recordings.

Two iterations of a similar experiment are performed using each feature. In each iteration, a single hidden layer multi-layer perception is used with 512 neurons in the hidden layer for training. We have experimented with the model, optimizing it for accuracy using threefold cross-validation. For the Chroma feature, the model performed ordinarily by achieving an average of 66% accuracy over ten repetitions, while the model tops 85% and an average of 81% using Swaragram using the same settings. We understand that the experiment performed by us is not enough to get the quantitative evidence, but it does indicate the usefulness of it in Indian music analysis. We will publish the results by thoroughly experimenting on the whole dataset and compare it with the Chroma feature's performance.

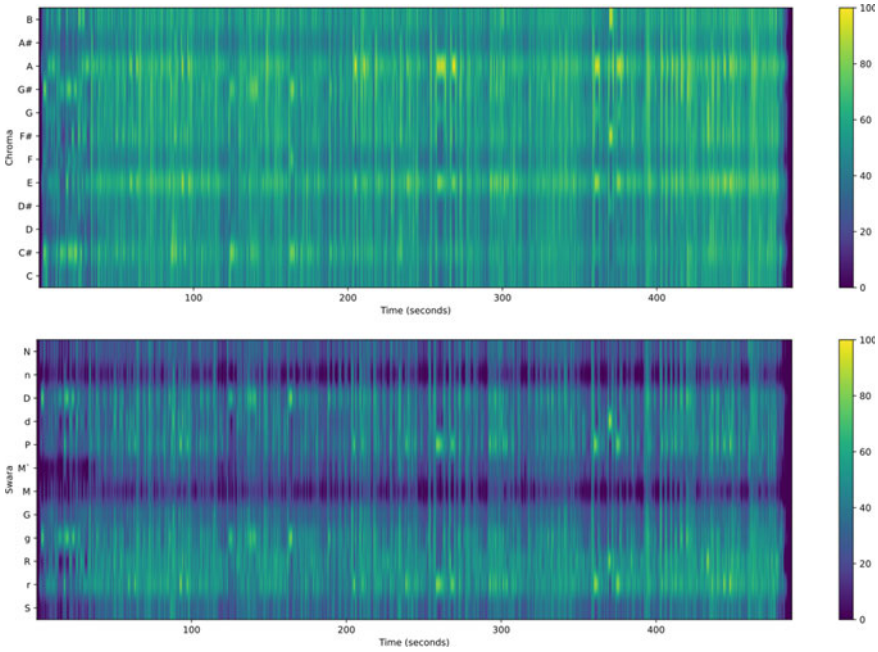


Fig. 1 Chromagram and Swaragram plot of ICM audio sample

We show Chroma and Swaragram plots of Indian classical music performance in Fig. 1. The audio sample was a vocal performance of raga Bhoopali [27] tuned at $A_{\#}4$ with a Tanpura. Raga is the melodic structure in Indian classical music having its grammar like Pakad, Vadi (king note), Samvadi (queen note), etc.—see Table 4—which guides performers. The upper plot in Fig. 1 shows the Chroma feature representation of our selected music sample. We can observe lots of bright areas almost all over. These bright areas represent the presence of corresponding notes at any given interval of time.

The grammar of raga Bhoopali is given in Table 4. It depicts that some of the swaras are forbidden in performance. The forbidden swaras are present (around A

Table 4 Grammar of raga bhoopali

Swaras	S R G P D (M and N are forbidden)
Aaroh-Avaroh	S R G P D S' - S' D P G R S R, D S
Jati	Audhav-Audhav
Thaat	Kalyan
Vadi/Samvadi	G/D
Pakad	RDS, RPG, PDS, SDP, GDP, GRS

and E, if we take Shrutis on the Western equal-tempered scale) in the Chroma feature representation, we see almost all areas lit. If we look at the Swaragram of our music sample shown in the lower plot of Fig. 1, we observe the absence of M and N swaras, which are forbidden in this raga. It can be argued that raga Bhoopali only employs Shudha swaras, and our plots also show the presence of Komal swaras. This is due to the ornamentation used by Indian music performers, where they glide from one swaras to another as given by the Chalan of that raga. The Swaragram feature can be further processed to make it tolerant of tonal and temporal variations, and we encourage others to do so.

5 Conclusions and Future Work

In this article, we have presented a feature inspired by the Chroma that is geared toward Indian music. The feature is also implemented as a Python module along with other utility functions. The Python module is released under an open license for encouraging other researchers to contribute to MIR with Indian music in focus. In addition, we have presented an application of this tool for raga identification in Indian classical music.

We suggest others to improve the feature technique and Python module as a whole by applying smoothing techniques like normalization, etc., to have better tolerance over noise and loudness variations, or adding other functionality to it. Another suggested improvement is integrating the automation of this tool with tonic identification into a seamless tool.

Acknowledgements This research was funded under grant number: ECR/2018/000204 by the Science & Engineering Research Board (SERB).

References

1. Singh Y (2020) Swaragram. https://github.com/yeshwantsingh/swaragram. Last accessed 2 May 2020
2. Gudmundsdottir HR (2010) Advances in music-reading research. Music Educ Res 12(4). https://doi.org/10.1080/14613808.2010.504809
3. Muller M, Ewert S (2011) Chroma toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In: Proceedings of the 12th international conference on music information retrieval (ISMIR)
4. Muller M (2007) Information retrieval for music and motion, vol 2. Springer, Heidelberg
5. Meredith D (2016) Computational music analysis. Springer, Heidelberg. https://doi.org/10.1007/978-3-319-25931-4
6. Meinard M (2015) Fundamental of music processing. Springer. https://doi.org/10.1007/978-3-319-21945-5
7. Dutta AK et al (2017) Signal analysis of Hindustani classical music. Springer, Singapore. https://doi.org/10.1007/978-981-10-3959-1

8. Bhatkhande (1995) Hindustani Sangeet Paddhati (Marathi/Hindi), vol 4. Popular Prakashan, Mumbai
9. Thakur DS (2015) The notion of twenty-two shrutis. Resonance 20(6):515–531
10. Burns EM (1999) Intervals, scales, and tuning. In The psychology of music. Academic Press, pp 215–264
11. Korzeniowski F, Widmer G (2016) Feature learning for chord recognition: the deep chroma extractor. ArXiv preprint arXiv:1612.05065
12. Ewert S, Muller M, Grosche P (2009) High resolution audio synchronization using chroma onset features. In: IEEE international conference on acoustics, speech and signal processing. IEEE, pp 1869–1872
13. Serra J, Gómez E, Herrera P, Serra X (2008) Chroma binary similarity and local alignment applied to cover song identification. IEEE Trans Audio, Speech, and Language Proc 16(6):1138–1151
14. Muller M, Kurth F, Clausen M (2005) Chroma-based statistical audio features for audio matching. In: IEEE workshop on applications of signal processing to audio and acoustics, IEEE, pp 275–278
15. Peeters G (2006) Chroma-based estimation of musical key from audio-signal analysis. ISMIR, pp 115–120
16. Bartsch MA, Wakefield GH (2005) Audio thumbnailing of popular music using chroma-based representations. IEEE Trans Multimedia 7(1):96–104
17. Salamon J, Gulati S, Serra X (2012) A multipitch approach to tonic identification in indian classical music. In: Proceedings of the 13th international society for music information retrieval conference (ISMIR)
18. Gulati S, Salamon J, Serra X (2012) A two-stage approach for tonic identification in Indian art music. In: Proceedings of the 2nd compMusic workshop; 2012 Jul 12-13; Istanbul, Turkey. Barcelona: Universitat Pompeu Fabra, pp 119–127
19. Harrison D (1994) Harmonic function in chromatic music: a renewed dualist theory and an account of its precedents. University of Chicago Press
20. LibROSA: Python package for music and audio analysis. https://librosa.github.io. Last accessed 22 March 2020
21. Essentia: Open-source library and tools for audio and music analysis, description and synthesis. https://essentia.upf.edu/. Last accessed 22 March 2020
22. Raga. https://en.wikipedia.org/wiki/Raga. Last accessed 2 May 2020
23. Gulati S, Serrà J, Ganguli KK, Şentürk S, Serra X (2016) Time-delayed melody surfaces for rāga recognition. In: Proceeding of the international society for music information retrieval conference (ISMIR). New York, USA, pp 751–757
24. Gulati S, Serrà J, Ishwar V, Şentürk S, Serra X (2016) Phrase-based Rāga recognition using vector space modeling. IEEE international conference on acoustics, speech and signal processing (ICASSP). Shanghai, China, pp 66–70
25. Chordia P, Rae A (2007) Raag recognition using pitch-class and pitch-class dyad distributions. In: Proceeding of international society for music information retrieval conference, pp 431–436
26. Dunya (2020) Music corpora and related software tools that have been developed as part of the CompMusic project. https://dunya.compmusic.upf.edu/. Last accessed 22 March 2020
27. Bhoopali R (2020) https://en.wikipedia.org/wiki/Bhoopali. Last accessed 2 May 2020

Machine Learning Approach for Audio Surveillance Using R

Nida Farheen and Subarna Chatterjee

Abstract With the rapid development of artificial intelligence, issues of public and personal safety are becoming increasingly prominent. Hence, surveillance-based security systems have attracted widespread attention and are being incorporated everywhere for secure and safer living options. People pay more and more attention to the surveillance systems because of their convenient and information-rich features. Video surveillance (VS) is very popular, but in some circumstances, the visual details are not decisive, whereas the use of audio can improve the reliability of surveillance systems Foggia (IEEE Trans Intell Trans Syst 17(1):279–288, 2016 [1]). Audio surveillance (AS) can be used as a biometric identity and security alert. In this work, a machine learning-based audio surveillance system is designed and simulated with R. It has two parts, the first is to identify the speaker and the second is to identify the gender of the speaker by using highly accurate, efficient, and reliable machine learning algorithms. These two are the most important aspects of any audio surveillance system. The algorithm is tested on the Open Speech and Language Resources (Open SLR) dataset. This dataset consists of both male and female audio samples. Here, three classifiers are used and compared.

Keywords Audio surveillance · R tool · Machine learning

1 Introduction

The details and changing behavior of surroundings can be captured and informed by audio surveillance systems. Compared to VS, the main advantages of AS systems are (1) they do not have to deal with variations in illumination conditions and can work during day and night [1] and (2) they require less storage space and therefore need less computation during information processing. But in open environments, the main

N. Farheen
MLIS, Ramaiah University of Applied Sciences, Bangalore, India

S. Chatterjee (✉)
Department of CSE, Ramaiah University of Applied Sciences, Bangalore, India
e-mail: subarna.cs.et@msruas.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_11

challenge is to eliminate the background noise. Tirumala et al. have identified feature extraction techniques as one of the most important aspects of speaker identification [2], which influences this process and performance. A review was conducted to recollect and survey various feature extraction methods and algorithms of speaker identification. The most successful algorithms are depending on the application areas. Requirements used for audio surveillance are given below:

Voice activity detection (VAD) [2, 3] is used to identify the presence of human speech.

Context recognition [3] algorithm is used to recognize the physical environment around the device to identify relevant sounds.

Filtering techniques, Kalman [10] and Weiner filters, are commonly used to denoise any noisy signals and predict future values based on the noise measurements.

Acoustic event detection (AES) method acts as an acoustic threshold for detecting sound.

Adaptive filtering techniques, IIR, FIR filters, and least mean square, are used in adaptive noise cancellation, adaptive equalization, and echo cancellation techniques.

Each algorithm works differently in different application areas. But generally, filtering, VAD, and AES algorithms work best since they extract the best features of the audio signal and can remove background noise.

The organization of the paper is as follows: Sect. 1 engraves the introduction, Sect. 2 presents a review of existing methods for speech recognition. Section 3 describes the methodology of the proposed algorithm. Section 4 covers the simulation results with discussion. Section 5 contains the concluding remarks.

2 Review of Existing Speech Recognition Methods

2.1 Feature Extraction

The various techniques used for feature extraction extrapolated for this work are as follows:

Mel-Frequency Cepstral Coefficients (MFCC) This feature extraction approach is the most popular approach in speech recognition. MFCC is extensively used in audio classification, speaker identification, and emotion recognition because of its good performance. The functional block for MFCC's is shown in Fig. 1.

The signals are cut into frames. For each frame, the periodogram is calculated to estimate the power spectrum. The features are subjected to the Mel-scale, which relates the perceived frequency to its actual measured frequency. The Mel-filter bank is applied to get the sum total energy in each filter. Then, the log is taken of the filter bank energies. The discrete cosine transform (DCT) is applied to all filter banks to decorrelate them. Only the DCT coefficients are kept, and rest are discarded.

Linear Predictive Cepstrum Coefficients (LPCC) is a technique that works on the principle of linear prediction analysis to extract the most important features of a speech signal. The coefficients altogether form a spectrum of features. The features extracted from the spectra use the energy values of the linearly arranged filter banks and put equal emphasis on the contribution of all the frequencies.

Information Extraction: This technique is an effective and required technique in speech and natural language processing (NLP). It is a feature extractor from audio voice signal, since it only extracts the important and most repeated words from the audio. It contains all the information in its database, such as linguistic knowledge of grammar, phonemes, and other semantics.

Lexicon plays a pivotal role in speech recognition for emotion analysis. It is a technique that creates a word cache of the vocabulary of languages, and contains all the linguistic catalogues of grammar, semantics, and morphological and phonological features of all different languages. It basically acts like a dictionary to the model for audio processing and makes use of it whenever required. It creates a link between the acoustic representation of a speech signal and the word sequence output from the recognizer, which in this case are various emotions.

Multilingual and Multimedia Content Extractor (MMCE) can be used in both speech and image processing. It is nothing but a content extraction pipeline to extract information like audio, text, video, image, hand/eye gestures, etc., from the input data. The raw input data is preprocessed, and all the important linguistic information is recognized, processed, classified, and interpreted.

Vector Quantization (VQ) [1] is the technique of mapping the feature vectors from a large feature vector space to a finite region of feature vector space. It works on the principle of block coding, where each region is called a cluster and is represented as a code word.

Convolutional Neural Network (CNN) [1]-based feature extraction and classification algorithms are very effective and accurate. They provide end-to-end encryption of the text. The densely connected neural nets, along with their hidden layers, transform the sequence of features extracted to the sequence of probability density distribution over the target of emotion labels.

Fig. 1 Functional block for MFCC extraction

2.2 Classification

The analysis of a feature classification techniques is given as follows:

Gaussian Mixture Model (GMM) [1, 5, 11] is a feature classifier that works on the principle of Naive Bayes' theorem of maximum likelihood. It works on the underlying probability distribution functions of the extracted features. The features extracted are classified based on their parametric probability density function (PDF).

Expectation Maximum (EM): This is a unique technique that finds the optimal values of the parameters chosen. It is flexible and can be used on GMM and HMM set of observations. It is based on the maximum likelihood principle and is an iterative process, which has two steps: E-step and M-step.

k-Nearest Neighbor (kNN) [12] performs classification and regression based on the proximity of nearest neighbors. This algorithm estimates the coefficients of the features extracted and takes the weighted average of the k-nearest neighbor. It performs classification of the input speech signal by the voting scheme of the nearest vectors available. It groups the closest vectors into the same cluster based on similarity measures calculated by statistics.

Hidden Markov Model (HMM) [1, 5, 11] is most extensively used in speech processing for emotion analysis. It is a simple technique that is based on the principles of the Markov chain model. It also takes probability distribution into account. HMM is feasible and easy to compute with; therefore, they are widely used for classifying the emotions in a speech signal. This method splits the entire speech signal into smaller samples, and these samples represent the hidden states of the model.

Support Vector Machine (SVM) [1, 11, 12] uses a likelihood ratio to classify the speech signals into desired classes labels of emotions. It constructs a hyperplane that separates the extracted feature vectors from the input speech signal and classifies them into similar classes of emotions. The feature vectors can be mapped from high input d-dimensional feature space to high d-dimensional output feature space.

Decision Trees or Random Forest does both classification and regression. It is a supervised form of learning, wherein the algorithm creates a set of forests with a number of trees having branches and nodes. It works on the principle of the decision trees algorithm. The decision trees pertain to a set of rules created by them after carefully observing the features of the parameters extracted. Depending on the features extracted, the trees of emotions are created and classified accordingly.

Statistical Analysis: These include various hypothesis testing methods, such as z-test, t-test, f-test, and p-test statistics. These hypothesis tests assume the probability values, called the critical values of the tests, and based on that either accept or reject the null hypothesis. Analysis of variance—ANOVA—is another technique, which calculates the variance of extracted features to inform the classifier which features to consider and which to discard. Mean vector and covariance matrix is another method in statistics that is used to calculate the mean of the feature vectors

and find the variance of the same in order to know their significance in the model. Linear regression is a statistical machine learning technique that classifies the features extracted from speech signals linearly into a model, separated by a line. It works on the principle of the slope of a straight line.

Artificial Neural Networks (ANN) [1]: This is an unsupervised machine learning algorithm that uses the perceptron to train and test the data. It has the ability to learn based on previous experience. The major role an ANN plays in speech recognition is the creation of its own powerful pattern recognition technique by learning from the dataset. The advantages are that it can compute in parallel and can handle complexities in data. It develops the system through learning and not programming.

Bayesian Classifier: In speech processing for emotion recognition, this algorithm works on the basis of word frequencies represented as the features. The classification is done based on the probability of occurrence of the words. The Bayesian classifier is highly scalable and uses likelihood parameters for classifying.

2.3 DataSet

The real-time audio samples were taken from an open Web source called OpenSLR (openslr.org). In that, a file called SLR41 was chosen, which has high-quality TTS data for Javanese speech. It contained multiple speakers, and both male and female audio samples were chosen.

3 Methodology

Two algorithms are designed for this problem. The first is to **identify the speaker**, and the second is to **identify the gender** of the speaker. The functional block of the proposed algorithm is shown in Fig. 2 and has five parts:- preprocessing of the audio samples, feature extraction, data preparation, classification, and the last block is used to train and test the samples for accurate predictions.

Fig. 2 Functional block of the proposed algorithm

3.1 Algorithm for Speaker Identification

STEP 1: Read audio files

The audio samples are read from the audio file location using the command "list.files" to access the files directly from their locations, and "readWave" to read each sample from the audio file. This is done for all five datasets.

STEP 2: Data preprocessing

Each audio sample is passed through a windowing technique. The Hamming window is used in this algorithm as it is used widely in audio surveillance. The hamming window compresses and reduces the speech signal without the loss of information. The hamming window has restricted the audio sample from the datasets to 512 points or samples in the audio wave.

STEP 3: Feature extraction

The acoustat function [15] is used to extract unique features of the audio wave. It works on the principle of short-term Fourier transform (STFT) that converts the audio wave into a frequency and time domain matrix. The resulting matrix will have a time and a frequency curve. Each curve is considered as a probability mass function (PMF) and is transformed into a cumulated distribution function (CDF).

From each **CDF**, the median (M), initial-percentile (P1) value, terminal-percentile (P2) value, and inter-percentile-range (IPR) are measured. P1, P2, and IPR are defined using a fraction parameter. The curve amplitude values are calculated in initial and terminal percentile. The function returns a list with ten items:

timecurve: This curve is a two-column matrix (x, y), the first column is time (s), and the second column the amplitude PMF.

freqcurve: This curve is a two-column matrix, the first column is frequency (kHz), and the second column is amplitude PMF.

timeP1 (initial-percentile), **timeM** (median), **timeP2** (terminal percentile), and **timeIPR** (inter-percentile range) are calculated with respect to time.

freqP1 (initial-percentile), **freqM** (median), **freqP2** (terminal-percentile), and **freqIPR** (inter-percentile range) are calculated with respect to frequency.
Therefore, a total of eight features are extracted from each sample of all five audio wave datasets.

STEP 4: Data preparation

After all eight important features are extracted from each of the dataset, the result that we get from 50 samples of four datasets having eight features each, and 12 samples from the fifth dataset having eight samples each, gives a total of 50 * 4 = 200 + 12 = 212 samples.

The total dimension of all the datasets are 212 * 8, i.e., 212 audio samples and eight extracted features. All the datasets are added row wise by using the command "rbind".

The independent output variable is added as the last column, and the classes are assigned labels. Therefore, including the output variable, the total dimension of the whole dataset becomes- 212 * 9.

The class label assigned to **Dataset 1** is **class 1**. Similarly, **Dataset 2, 3, 4** and **5** are assigned to **class 2, 3, 4** and **5**. These class labels define the classes to which the datasets belong when the model is trained and classified. The classifier, when trained and tested, will classify the dataset appropriately into its class labels to identify the correct speaker.

STEP 5: Classification

Three classification techniques are used in this algorithm. They are classification and regression, kNN, and random forest. The whole dataset is partitioned into two with ratio 7:3 which means out of the whole dataset, 70 samples are taken to train the model and the remaining 30 are taken for testing the model accuracy.

(a) Training: 70% of the data, i.e., out of 212 samples, 149 samples are used for training the classifiers. Their performance accuracy is checked by statistical summary.

(b) Testing: If the accuracy is good, then 30% of the data is used for testing. Otherwise, there must be some modeling issue with the classifier, and hence, the entire model needs to be changed.

The same steps must be repeated for the gender identification algorithm with the only difference being that there are only two classes to be classified into either male or female.

3.2 Algorithm for Gender Identification

STEP 1 to 3: Same as speaker identification.

STEP 4: Data preparation Same as speaker identification.
The datasets 1, 2, and 3 have male voice, and datasets 4 and 5 have female voices; hence, there can be only two classes for gender identification. The **datasets 1, 2, 3** are assigned to class label **1**, and **datasets 4, 5** are assigned to class level **2**. These class labels define the classes to which the dataset belong when the model is trained and classified. The model, when trained and tested for classification, classifies the dataset in either class 1 male or class 2 female.

STEP 5: Same as speaker identification.

4 Analysis and Results

The audio surveillance system is implemented on the audio samples of Javanese; a native Indonesian language taken from the Web open source OpenSLR. Five datasets are considered out of which the first four datasets have 50 samples each, and the fifth dataset has 12 samples of audio wave. The datasets are numbered orderly, where datasets 1, 2, 3 all have male voice audio. Datasets 4 and 5 have female voice audio waves to bifurcate gender classification. The following figures are the visual contents of the aforementioned algorithmic analysis and results.

Figures 3, 4, and 5 depict the time and frequency curve of the signal sample of dataset1, dataset2, dataset3, dataset4, and dataset5 using R tool. Table 1 shows the statistical summary of the five samples from the five datasets. All the wave object have the same sampling rate 48000 Hz, all channels are Mono, and all have the same bit value, which is 16. The number of samples present in each wave files and the duration of the waves are different.

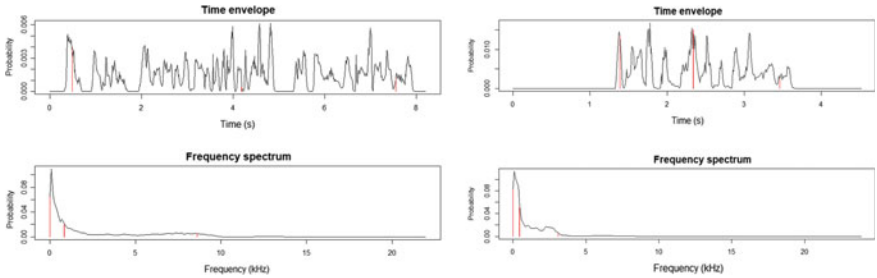


Fig. 3 Time and frequency curve of a signal sample of dataset1 and dataset2

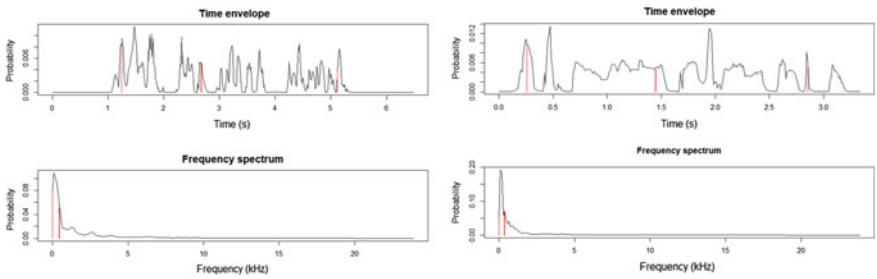


Fig. 4 Time and frequency curve of a signal sample of dataset3 and dataset4

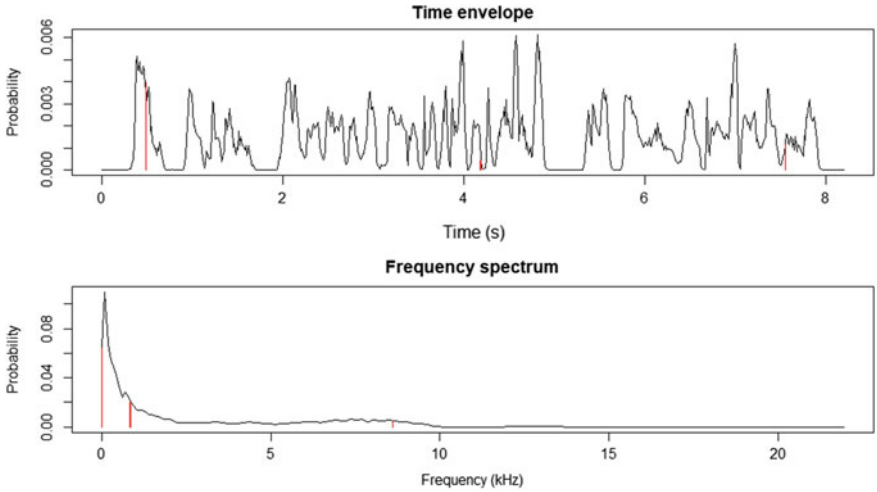


Fig. 5 Time and frequency curve of a signal sample of dataset5

Table 1 Statistical summary of the samples

	No. of samples	Duration (s)	Sampling rate (Hz)	Channels (Mono)	PCM (int)	Bits
Wave object 1	187,676	3.91	48,000	Mono	True	16
Wave object 2	172,885	3.6	48,000	Mono	True	16
Wave object 3	189,000	3.94	48,000	Mono	True	16
Wave object 4	166,800	3.48	48,000	Mono	True	16
Wave object 5	188,173	3.92	48,000	Mono	True	16
Summary statistics for channel(s):						
	Min.	First Qu.	Median	Mean	Third Qu.	Max.
Wave object 1	−24,978.000	−582.000	26.000	0.227	1113.000	25,822.000
Wave object 2	−26,387.000	−469.000	26.000	1.523	997.000	17,834.000
Wave object 3	−26,404.000	−220.000	4.000	1.437	540.000	19,842.000
Wave object 4	−29101.000	−401.000	10.000	2.319	1065.000	16,622.000
Wave object 5	−26,905.000	−375.000	62.000	3.093	811.000	23,910.000

The classifier model is tested with audio samples, and performance of each model is validated with a confusion matrix and statistical measurements like mean, median, kappa values, and ROC curves. Table 2 shows the confusion matrix, statistical summary, and testing results of classifier CART, which are measured using R tool. The confusion matrix, statistical summary, and testing results for three machine learning classifiers are given in Table 2 for CART, Table 3 for kNN, and Table 4 for random forest.

Table 2 Confusion matrix, statistical summary, testing results of classifier cart. From R tool

Confusion matrix and statistics					
	Reference				
Prediction	1	2	3	4	5
1	15	0	0	0	0
2	0	15	0	0	0
3	0	0	15	0	1
4	0	0	0	15	0
5	0	0	0	0	2
Overall statistics					
Accuracy:			0.9841		
95% CI:			(0.9147, 0.9996)		
No information rate:			0.2381		
P-value [Acc > NIR]:			<2.2e−16		
Kappa:			0.9793		
Mcnemar's test P-value			NA		
Statistics by class:					
	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	1.0000	1.0000	1.0000	1.0000	0.66667
Specificity	1.0000	1.00000	0.9792	1.0000	1.00000
Pos pred value	1.0000	1.0000	0.9375	1.0000	1.00000
Neg pred value	1.0000	1.0000	1.0000	1.0000	0.98361
Prevalence	0.2381	0.2381	0.2381	0.2381	0.04762
Detection rate	0.2381	0.2381	0.2381	0.2381	0.03175
Detection prevalence	0.2381	0.2381	0.2540	0.2381	0.03175
Balanced accuracy	1.0000	1.0000	0.9896	1.0000	0.83333

Table 5 shows the comparison of the classifiers CART, kNN, and random forest. Accuracy and kappa values are calculated here for comparison. Table 5 displays the best results for the random forest classifier. The accuracy is obtained from the confusion matrix for all the three classifiers. The kappa coefficient (K) [14] is a statistical value, which measures inter-rater reliability for qualitative items. Intra-rater reliability is related to the different calculations done by the same person. Both accuracy and kappa should be higher with better measurements. The average accuracy of the algorithm can be seen in Table 5: kNN is 93.3%, CART is 98.9%, and random forest is 100%. The average kappa of the algorithm can be seen in Table 5: kNN is 91.3%, CART is 98.6%, and random forest is 100%.

Table 3 Confusion matrix, statistical summary, and testing results of classifier kNN from R tool

Confusion matrix and statistics					
	Reference				
Prediction	1	2	3	4	5
1	15	0	0	0	0
P	0	14	1	0	0
P	0	0	14	0	1
P	0	0	0	15	0
5	0	0	0	0	2
Overall statistics					
Accuracy :		0.9524			
95% CI :		(0.8671, 0.9901)			
No information rate:		0.2381			
P-value [Acc > NIR]:		<2.2e−16			
Kappa:		0.938			
Mcnemar's test P-value		NA			
Statistics by class:					
	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	1.0000	0.9333	0.9333	1.0000	0.66667
Specificity	1.0000	0.9792	0.9583	1.0000	1.00000
Pos pred value	1.0000	0.9333	0.8750	1.0000	1.00000
Neg pred value	1.0000	0.9792	0.9787	1.0000	0.98361
Prevalence	0.2381	0.2381	0.2381	0.2381	0.04762
Detection rate	0.2381	0.2222	0.2222	0.2381	0.03175
Detection prevalence	0.2381	0.2381	0.2540	0.2381	0.03175
Balanced accuracy	1.0000	1.0000	0.9896	1.0000	0.83333

5 Conclusion

The algorithms are implemented using the acoustat function and STFT algorithm for audio surveillance. The classifiers used are kNN, CART, and random forest. The validation and analysis for each of the classifier is compared by measuring the accuracy and statistic measurements such as mean, median, and kappa values. The average accuracy of the algorithm kNN is 93.3%, CART is 98.9%, and random forest is 100%. The average kappa of the algorithms kNN is 91.3% , CART is 98.6%, and random forest is 100%.

Table 4 Confusion matrix, statistical summary, and testing results of classifier random forest classifier from R tool

Confusion matrix and statistics					
	Reference				
Prediction	1	2	3	4	5
1	15	0	0	0	0
2	0	15	0	0	0
3	0	0	15	0	1
4	0	0	0	15	0
5	0	0	0	0	2
Overall statistics					
Accuracy:			0.9841		
95% CI:			(0.9147,0.9996)		
No information rate:			0.2381		
P-value [Acc > NIR]:			<2.2e−16		
Kappa:			0.9793		
Mcnemar's test P-value			NA		
Statistics by class:					
	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	1.0000	1.0000	1.0000	1.0000	0.66667
Specificity	1.0000	1.00000	0.9792	1.0000	1.00000
Pos pred value	1.0000	1.0000	0.9375	1.0000	1.00000
Neg pred value	1.0000	1.0000	1.0000	1.0000	0.98361
Prevalence	0.2381	0.2381	0.2381	0.2381	0.04762
Detection rate	0.2381	0.2381	0.2381	0.2381	0.03175
Detection prevalence	0.2381	0.2381	0.2540	0.2381	0.03175
Balanced accuracy	1.0000	1.0000	0.9896	1.0000	0.83333

Table 5 Statistical accuracy results of classifiers CART, kNN, and random forest, snippet of code from R tool

Models: kNN, CART, random forest						
Number of resamples: 30						
	Min.	First Qu.	Median	Mean	Third Qu.	Max.
Accuracy						
kNN	0.8000	0.9230769	0.9375	0.9408663	1	1
CART	0.9375	1.0000000	1.0000	0.9979167	1	1
Random forest	1.0000	1.0000000	1.0000	1.0000000	1	1
Kappa						
kNN	0.7398844	0.9003788	0.919998	0.9233482	1	1
CART	0.9191919	1.0000000	1.000000	0.9973064	1	1
Random forest	1.0000000	1.0000000	1.000000	1.0000000	1	1

References

1. Foggia P, Petkov N, Saggese A, Strisciuglio N, Vento M (2016) Audio surveillance of roads: a system for detecting anomalous sounds. IEEE Trans on Intell Trans Syst 17(1):279–288
2. Tirumala SS, Shahamiri SR, Garhwal AS, Wang R (2017) Speaker identification features extraction methods: a systematic review. Expert Syst Appl 90:250–271
3. Ntalampiras S (2012) Audio surveillance in critical infrastructure security: assessment. Detection, Response, WIT Press, Prevention, pp 191–205
4. Harma A, McKinney MF, Skowronek J (2005) Automatic surveillance of the acoustic activity in our living environment. In: IEEE international conference on multi-media and expo, Amsterdam (2005)
5. Dufaux A, Bezacier L, Ansorge M, Pellandini F (2000) Automatic sound detection and recognition for noisy environment. In: European signal processing conference, Finland, pp 1033–1036
6. Rabaoui A, Davy M, Rossignol S, Lachiri Z, Ellouze N (2007) Using one-class SVMs and wavelets for audio surveillance systems. IEEE Trans Inf Forensic Sec
7. Morton KD (2010) Bayesian techniques for adaptive acoustic surveillance, Ph.D. Thesis, Duke University
8. Radhakrishnan R, Divakaran A, Smaragdis P (2005) Audio analysis for surveillance applications. In: Applications of signal processing to audio and acoustics. IEEE Workshop on, pp 158–161
9. Godsill SJ, Rayner PJW (1998) Digital audio restoration -a statistical model based approach, book, vol 1
10. Zwiecki MN, Ciołek M, Cisowski K (2015) Elimination of impulsive disturbances from stereo audio recordings using vector autoregressive modeling and variable-order Kalman filtering. IEEE Trans Audio Speech Lang Process 23(6):970–981
11. Crocco M, Cristani M, Trucco A, Murino V (216) Audio surveillance: a systematic review. ACM Comput Surv 48(4):52:1–52:46
12. Nannapaneni R, Chatterjee S (2020) Human emotion recognition through facial expressions. MISS2k20. Springer algorithms for intelligent systems (AIS) Series, ISSN 2524-7565
13. Arslan Y, Canbolat H (2017) A sound database development for environmental sound recognition. In: Signal processing and communications applications conference (SIU)
14. Zaiontz C Cohen's Kappa. http://www.real-statistics.com/reliability/interrater-reliability/cohens-kappa/
15. Fristrup KM, Watkins WA (1992) Characterizing acoustic features of marine animal sounds. Woods Hole Oceanographic Institution Technical Report WHOI-92-04

An Artificial Intelligence-Based Approach Towards Segregation of Folk Songs

Himadri Mukherjee, Ankita Dhar, Sk. Md. Obaidullah, Santanu Phadikar, Kaushik Roy, and Sukumar Nandi

Abstract Music genre/type classification has been an active area of interest for researchers. There has been development of disparate systems for Western music, but a scarcity of such systems is observed for Indian music especially the folk songs. The folk songs of different regions bring out the cultural heritage and richness of those places. Bengal encompasses a diverse and large variety of folk songs which is very rich both musically and culturally. There is a pressing need for automatic systems to categorize such songs to make them easily accessible and thereby preserve our culture. This is very much important because it is hard for a person with little knowledge to distinguish the folk song types and thus dive deeper into them. To the best of our knowledge, there are no currently available systems to distinguish folk songs, especially of Bengal. Here, an artificial intelligence-based system is presented for this task which distinguishes the five major types of folk songs of Bengal. Experiments were performed with over 43 K clips, and the best correct classification rate of 99.73% was obtained with LPCC-based features considering extremely short length clips.

Keywords Folk songs · LPCC · Ensemble learning · Artificial intelligence

H. Mukherjee (✉) · A. Dhar · K. Roy
Department of Computer Science, West Bengal State University, Kolkata, India

Sk. Md. Obaidullah
Department of Computer Science and Engineering, Aliah University, Kolkata, India

S. Phadikar
Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata, India

S. Nandi
Department of Computer Science and Engineering, Indian Institute of Technology, Guwahati, India
e-mail: sukumar@iitg.ernet.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_12

1 Introduction

Analysis of songs has been an interesting field for the researchers for quite some time. The development of artificial intelligence has led to the emergence of different avenues including song categorization, automatic song recommendation and song identification. Songs can be categorized into different genres namely rock, jazz, blues and folk to name a few. Among these, folk songs form an extremely versatile group. There has been development of several systems for songs of other genres, but such advances have not found its way in the genre of folk songs. Among the folk songs of different countries, Indian folk songs are extremely diverse and present the cultural variety of the country. Folk songs exhibit disparate characteristics based on regions of origin which also play a part in conveying the culture of that place. A wide variety of folk songs can be found in Bengal in the different districts. It is very much important to preserve them as this in turn will preserve the culture of the place. Primarily, an automatic categorization system is needed, which can segregate the folk songs based on their types and thus make them easily accessible. This is essential because it is extremely hard for people with no previous musical knowledge to distinguish the type of folk songs. There has been development of systems to distinguish song covers, but unfortunately none of those systems deal with folk songs to the best of our knowledge.

Chang et al. [1] used convolutional neural networks for distinguishing cover songs. Experiments were performed on the MIREX cover song dataset. It consists of 3300 covers and 496200 non-covers. The training set was composed of 2113 cover and non-cover pairs. Using their proposed approach, they obtained a value of 8.04 for the mean number of covers distinguished in the top-10. Serra and Gomez [2] used tonal descriptors for cover song identification. Experiments were performed on the MIREX 2007 data consisting of 30 covers each of which had 11 versions. A detailed account of the obtained results is given in [2]. Serra et al. [3] have discussed disparate approaches for identifying covers. They have talked about types of covers and different musical aspects. Different features have been discussed as well as structure, key and tempo invariance. A ray of light has been cast on evaluation measures and similarity computations as well.

Chang et al. [4] worked with AAC files for cover song identification. Experiments were performed with 12-dimensional chroma features on a dataset consisting of 12960 training instances and 5236 test instances. An accuracy of 62% was reported in the experiments. Ke et al. [5] proposed a spectrogram-based technique for identifying music. They trained their system with 78 songs which were played using low-quality speakers and reported a highest precision of 96%. Heo et al. [6] used distance-based feature for distinguishing cover songs. They experimented with the MIREX dataset where 1175 songs were used in the training phase. Several experiments were performed with different set-ups which are detailed in [6]. Janssen et al. [7] attempted to distinguish melodic sections in folk songs. They used similarity-based approach in the study. They compared different similarity-based approaches including city block distance, Euclidean distance, correlation distance, etc., to name a few on a dataset of

Dutch folk songs. Lee et al. [8] used song to song cross-similarity along with CNN for distinguishing cover songs. Experiments were performed on a dataset consisting of 3300 covers and 496200 non-covers, and an average precision of 93.18% was reported.

Panagiotou and Mitianoudis [9] used GMM and MFCC for identifying songs. They also introduced a novel implementation of PCA. Experiments were performed with 10000 songs, and a correct classification rate of 96% was reported for ordinary condition. Further experimentations were also performed with different scenarios. Wang et al. [10] used structural attributes for song identification. Experiments were performed with 10000 songs, and an accuracy of 99.7% was reported with clips of 5 s keeping the SNR value at 0db. Maddage [11] has discussed automatic identification of structure in music. He has talked about chord identification, singing voice boundary detection, song structure detection encompassing intro, verse and chorus to name a few. Furthermore, several experiments were performed for detection of different aspects with 50 songs whose results are detailed in [11].

It is observed that here has been development of disparate systems for analysis of songs in different avenues. However, there has hardly been any such development related to the folk songs of Bengal. There is a huge diversity in these folk songs, and some of them differ from each other very minutely. Hence, distinguishing them is a major challenge. Another challenge is the availability of datasets. These songs are not concentrated in a single place but rather distributed. It has been noted that there are disparate type of folk songs which have hardly been digitized. Hence, putting up a dataset of the folk songs of Bengal is a very difficult task. In this paper, an automatic technique is introduced for distinguishing Bangla folk songs with LPCC-based features.

2 Dataset

Data is a very important aspect of an experimental work. Both the quality and quantity have a significant role in the development of the models. It has to be made sure that a dataset encompasses real-world varieties. Presently, there is no standard dataset of folk songs of Bengal to the best of our knowledge. Hence, a dataset was put together encompassing five major types/folk regions namely Bhawaiya, Bhatiali, Baul, Jhumur and Kirtan [12]. The dataset comprised of studio quality data as well as open air condition data. The songs were sung by both males and females. In several songs, similar instruments were present. There was also presence of similar type of noises as often seen in real world. The audios were in .wav format (44.1 KHz) and stereo mode. The audios were not subjected to pre-processing for removal of ambient noises to ensure real-world conditions. Further, there was presence of different introductory music prior to beginning and end of a composition. Sometimes, there were short descriptions. There were also instrumental sections amidst vocal sections. Such attributes were not removed to test the system's performance and

Table 1 Number of clips in each of the engendered datasets with length of the clips in seconds

Dataset	Bhawaiya	Bhatiali	Baul	Jhumur	Kirtan	Total
D_1 (0.5)	9754	8569	7301	10357	7195	43,176
D_2 (1)	4878	4285	3651	5179	3598	21,591
D_3 (2)	2440	2143	1826	2590	1799	10,798
D_4 (3)	1627	1429	1217	1727	1200	7200
D_5 (4)	1220	1072	913	1295	900	5400
D_6 (5)	976	857	731	1036	720	4320
D_7 (6)	814	715	609	864	600	3602
D_8 (7)	698	613	522	740	514	3087
D_9 (8)	611	536	457	648	450	2702
D_{10} (9)	543	477	406	576	400	2402
D_{11} (10)	489	429	366	519	360	2163

robustness for uncleaned data. The collected songs were used for the generation of 11 datasets ($D_1 - D_{11}$) comprising 0.5 s, 1 s. 2 s and so on up to 10 s long clips. The generated datasets are detailed in Table 1.

3 Proposed Methodology

The audios were framed into smaller segments which were then subjected to windowing. This was followed by extraction of LPCC-based features. The extracted features were then used to distinguish the folk songs using an ensemble learning-based technique. The system is illustrated in Fig. 1.

3.1 Feature Extraction

3.1.1 Framing

The information contained in an audio clip shows high deviations all through. This creates problems during analysis. To address this issue, an audio clip is divided into

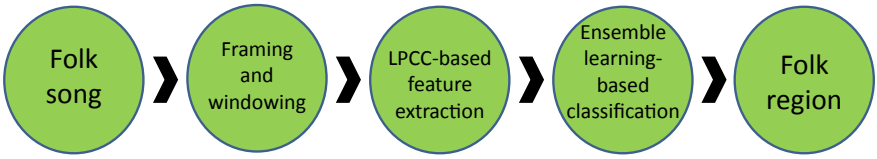


Fig. 1 Proposed methodology

small segments called frames. Here, the audios were divided into 256 sample point frames in an overlapping manner [13]. Each of the frames had a 100 sample overlap. The number of obtained (F)-sized frames (N) with an overlap of (L) for an audio having (n) samples is mathematically presented in Equation (1).

$$N = \left\lceil \frac{n - F}{L} + 1 \right\rceil \tag{1}$$

3.1.2 Windowing

Post framing it is often observed, that the starting and ending points of the sub signal (frame) are not aligned. This is a major problem in frequency-based analysis which produces misleading results. This is termed as spectral leakage. To ensure continuity, the frames are multiplied with a windowing function. Hamming window (W) was used in this experiment due to its suitability for audio signals [14]. It is mathematically illustrated in Equation (2) where $0 \leq f \leq 255$.

$$W(f) = 0.54 - 0.46 \cos\left(\frac{2\pi f}{F - 1}\right) \tag{2}$$

3.1.3 LPCC-Based Feature Generation

Linear predictive analysis was performed on each of the frames for obtaining the linear predictive cepstral coefficients (LPCCs) [13]. In this technique, a present signal sample is represented using its earlier samples. The xth sample of a signal (S) can be presented using the previous R samples as shown in Equation (3). Here, c_1, c_2, c_3,…c_R are the linear predictive coefficients (LPCs) or predictors.

$$S(x) \approx c_1 S(x-1) + c_2 S(x-2) + \cdots + c_R s(x-R) \tag{3}$$

The error ($e(x)$) involved in prediction of ($\hat{S}$(x)) (predicted sample) and actual sample $S(x)$ is presented in Equation (8). The total error is presented in Equation (5) which is minimized to form the coefficients where (n) presents the frame size.

$$e(x) = S(x) - \hat{S}(x) = S(x) - \sum_{i=1}^{R} c_i S(x-i) \tag{4}$$

$$E_x = \sum_{n} \left[S_x(n) - \sum_{i=1}^{R} c_i S_x(n-i) \right]^2 \tag{5}$$

The recursive technique to form the cepstral coefficients (C) is presented in Equation (6).

$$\left.\begin{aligned} C_0 &= \log_e R \\ C_n &= c_n + \sum_{i=1}^{n-1} \frac{i}{n} C_i s_{n-i}, \text{ for } 1 < n < R \text{ and} \\ C_n &= s_n + \sum_{i=n-p}^{n-1} \frac{i}{n} C_i s_{n-i}, \text{ for } n > R \end{aligned}\right\} \tag{6}$$

Audios are often of different lengths in reality. Since the features are extracted in frame level, hence features of disparate dimensions are generated. This poses difficulty in analysis. To handle this, the sum of the coefficients for each of the bands (frequency ranges) was computed. This sum was used to arrange the bands in an increasing order, and the band number sequence was extracted as a feature (F_1). The standard deviations for every band were tabulated which formed the second feature (F_2). A combined feature (F_3) was also obtained. This was generated by putting together F_1 and F_2. This not only produced features of uniform length but also the feature dimension was brought down as well. When only ten-dimensional LPCCs were extracted for an audio of 1 s, a feature of 4400 dimension was obtained (440 frames). However, using this technique, a feature of only 20 dimensions was produced (ten grades and ten standard deviations). In this work, 10, 20 and 30 dimensional LPCCs were extracted which produced even 20, 40 and 60 dimensional features which were invariant of the length of the audios. The feature trend for 40-dimensional F_3 features on D_2 (best performance) is shown in Fig. 2.

3.2 Ensemble Learning-Based Classification

The methodology involving combination of multitudinous learners to solve a problem is known as ensemble learning [15]. In this method, the multiple learners generate a set of rules during training whose decisions are combined to produce an output at the time of testing. This is different as compared to standard machine learning algorithms where only a single rule/hypothesis is constructed during training. Due to the technique of involving multiple rules, the generalization capacity of ensemble learners is higher.

A random forest-based classifier [15] was used here. Random forests are composed of a set of decision trees ($t_{1\ldots p}$). It is very much suitable for real-world problems because it can efficiently estimate missing data and uphold system performance in the absence of a large chunk of data [15]. Further, it is very much suitable for parallel processing systems as each of the decision trees can be trained independently. This makes it suitable for extremely large datasets. For each of the trees, a random vector ($\Theta_{1\ldots p}$) is generated with same distribution. However, Θ_k is independent of Θ_1 to Θ_{k-1}. The training set in addition to the Θ_k produces the kth classifier h (ip, Θ_k), where ip denotes the input vector. The margin function (MAR) for an ensemble (p) involving the random vector (Z) is presented as

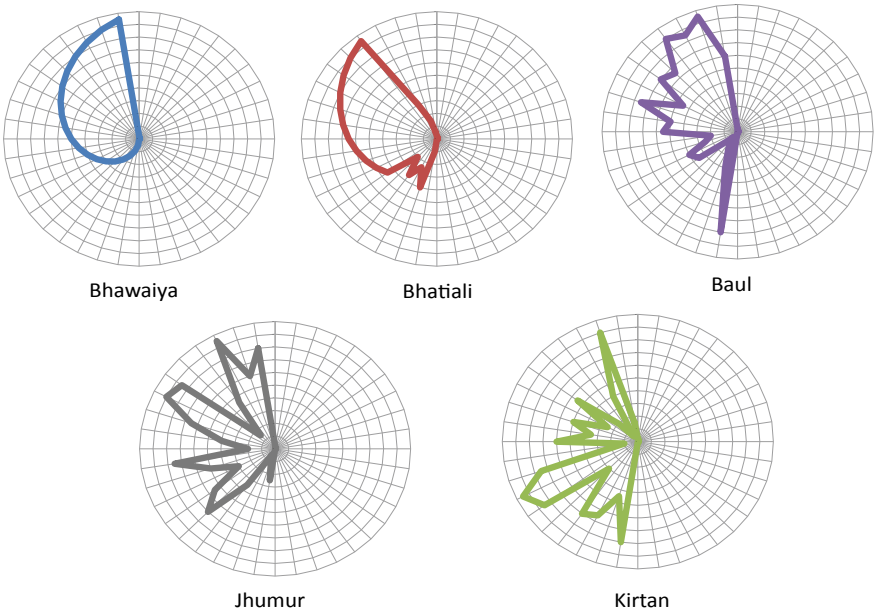


Fig. 2 Trend of 40-dimensional F_3 features on D_2

$$\mathrm{MAR}(X, Z) = av_p I(h_p(X) = Z) - \max_{j \neq Z} av_p I(h_p(X) = j) \tag{7}$$

where $h_p(X) = h(x, \Theta_p)$ and I corresponds to the indicator function. The generalization error and its convergence are shown in the following equations for a maximum random vectors.

$$PE^* = P_{X,Z}(mg(X, Z) < 0) \tag{8}$$

$$P_{X,Z}(P_\Theta(h(X, \Theta) = Z) - \max_{j \neq Z} P_\Theta(h(X, \Theta) = j) < 0) \tag{9}$$

4 Result and Discussion

Initially, the datasets $D_1 - D_{11}$ were subjected to the 10-, 20- and 30-dimensional grade and standard deviation features with fivefold cross-validation whose results are presented in Tables 2 and 3.

It is observed from the Tables that in the case of the grade features (F_1), the highest accuracy was obtained with 20-dimensional features on the dataset D_2. For the standard deviation features (F_2), the best result was obtained for the 8 s long clips with 30-dimensional features.

Table 2 Results for datasets $D_1 - D_{11}$ using 10-, 20- and 30-dimensional F_1 features

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}
10	98.62	98.95	94.16	91.44	98.09	95.21	95.72	93.75	89.86	95.09	87.75
20	98.84	**99.32**	98.14	95.22	95.5	96.25	95.06	94.01	97.71	85.1	93.02
30	99.3	99.3	97.18	95.53	97.67	96.5	93.84	94.95	98.22	85.89	89.37

Bold values indicates the best resluts

Table 3 Results for datasets $D_1 - D_{11}$ using 10-, 20- and 30-dimensional F_2 features

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}
10	67.96	73.58	78.3	80.92	82.04	83.66	83.29	83.67	84.05	84.22	84.7
20	73.87	74.65	83.89	85.39	86.89	87.69	87.73	88.44	88.38	88.55	88.35
30	76.31	81.72	85.54	88	88.61	88.84	89.51	90.28	**90.9**	89.34	90.25

Bold values indicates the best resluts

Table 4 Results for datasets $D_1 - D_{11}$ using 20-, 40- and 60-dimensional F_3 features

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	D_{11}
10	98.63	99.53	97.07	97.64	99.02	98.52	98.67	98.67	97.26	98	97.13
20	98.62	**99.69**	99.28	99.11	99.13	98.94	98.56	99.32	99.3	96.5	97.5
30	98.6	99.67	99.06	98.85	99.36	99.05	98.56	98.25	98.56	97.13	97.87

Bold values indicates the best resluts

Table 5 Inter-class confusions for D_2 using 40-dimensional F_3 features

	Bhawayia	Bhatiali	Baul	Jhumur	Kirtan
Bhawayia	4878	0	0	0	0
Bhatiali	45	4236	4	0	0
Baul	0	2	3647	0	2
Jhumur	0	0	2	5168	9
Kirtan	0	0	0	4	3594

Next, features (F_1) and (F_2) were merged (F_3) whose outcome is shown in Table 4. It is noted that the system accuracy increased to 99.69% which is higher than both (F_1) and (F_2) individually.

The inter-class confusions for this are tabulated in Table 5. It is noted that Bhawaiya songs were classified with 100% accuracy. The lowest accuracy was obtained for the Bhatiali songs. On detailed analysis, it was observed that there were several instances where similar instruments were played in some of the Bhawaiya as well as Bhatiali songs. Some of the recordings were in the outdoor condition as well. On analysing those audios, it was seen that there was a similarity of frequency content of the ambient noises which further led to this confusion. The best result for the 1 s long clips is because on increasing the clip length, more information was available to analyse, and in most cases, the similar portions were becoming the majority. Next, the folds of cross-validation were increased for F_3 as shown in Table 6.

Table 6 Results for different cross-validation folds on D_1 using 40-dimensional F_3 features

Folds	5	6	7	8	**9**	10
Accuracy (%)	99.69	99.69	99.71	99.71	**99.72**	99.68

Bold values indicates the best resluts

Table 7 Inter-class confusions for D_2 using 40-dimensional F_3 features and nine-fold cross-validation

	Bhawayia	Bhatiali	Baul	Jhumur	Kirtan
Bhawayia	4878	0	0	0	0
Bhatiali	44	4239	2	0	0
Baul	0	3	3647	0	1
Jhumur	0	0	2	5170	7
Kirtan	0	0	0	2	3596

Table 8 Results for training iterations on D_1 using 40-dimensional F_3 features and nine-fold cross-validation

Iteration	100	200	**300**	400	500	600
Accuracy (%)	99.72	99.71	**99.73**	99.72	99.72	99.72

Bold values indicates the best resluts

It is noted that nine-fold cross-validation produced the best performance whose confusion matrix is presented in Table 7. An increase of 0.03% was observed in the overall system performance as compared to the earlier set-up. In this case, the classification accuracy for the Kirtan and Jhumur songs improved. The same was observed for the Bhatiali songs as well. In this case, an increase of 0.07% in the recognition accuracy was observed over the fivefold scenario.

Finally, the training iterations were increased (initially 100) whose results are presented in Table 8. A minute increase in the system's performance was obtained for 300 iterations whose confusion matrix is presented in Table 9. In this case, the classification accuracy remained to be same for the Kirtan, Jhumur, Baul and Bhawaiya songs. In the case of the Bhatiali songs, a 0.5% increase in the classification accuracy was seen as compared to the 100 iteration set-ups.

The performance of a few other popular classifiers [16] namely multi-layer perceptron (MLP), support vector machine (SVM), bayesnet (BN), naïve bayes (NB), radial basis function (RBF), simple logistic (SL), liblinear (Lib) was also compared with the proposed approach whose results are presented in Fig. 3.

Table 9 Inter-class confusions for D_2 using 40-dimensional F_3 features, nine-fold cross-validation and 300 training iterations

	Bhawayia	Bhatiali	Baul	Jhumur	Kirtan
Bhawayia	4878	0	0	0	0
Bhatiali	42	4241	2	0	0
Baul	0	3	3647	0	1
Jhumur	0	0	2	5170	7
Kirtan	0	0	0	2	3596

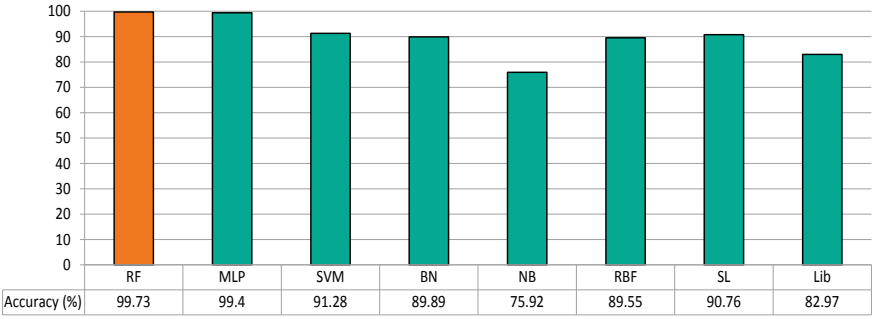


Fig. 3 Performance of different machine learning techniques

5 Conclusion

Here, a system is presented for distinguishing the major folk song types of Bengal. The experiments were performed with over 43K clips, and the best average recognition rate of 99.73% was obtained. This is very encouraging considering the diverse sources and ambiances of the clips. Moreover, the best result was obtained with clips of only 1 s, thereby pointing to the system's capability to handle clips of very short duration. In future, the dataset will be extended to encompass other folk song types of not only Bengal but also entire India. Experiments will also be performed with folk songs from other countries. Different spectral modelling approaches will be used in the future along with other machine learners for better performance. We also plan to apply deep learning to categorize the folk songs along with different pre-processing techniques to tackle the presence of disparate ambient noise sources. Efforts will also be made towards the separation of accompaniments from the voices for separate analysis and better decision-making.

References

1. Chang S, Lee J, Choe SK, Lee K (2017) Audio cover song identification using convolutional neural network. ArXiv preprint arXiv:1712.00166
2. Serra J, Gómez E (2007) A cover song identification system based on sequences of tonal descriptors. Music Inf Retrieval Eval exch (MIREX)
3. Serra J, Gómez E, Herrera P (2010) Audio cover song identification and similarity: background, approaches, evaluation, and beyond. Advances in music information retrieval. Springer, Berlin, Heidelberg, pp 307–332
4. Chang TM, Chen ET, Hsieh CB, Chang PC (2013) Cover song identification with direct chroma feature extraction from AAC files. In: 2013 IEEE 2nd global conference on consumer electronics (GCCE). IEEE, pp 55–56
5. Ke Y, Hoiem D, Sukthankar R (2005) Computer vision for music identification. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) Vol 1. IEEE, pp 597–604
6. Heo H, Kim HJ, Kim WS, Lee K (2017) Cover song identification with metric learning using distance as a feature. In: ISMIR pp 628–634
7. Janssen B, Van Kranenburg P, Volk A (2017) Finding occurrences of melodic segments in folk songs employing symbolic similarity measures. J New Music Res 46(2):118–134
8. Lee J, Chang S, Choe SK, Lee K (2018) Cover song identification using song-to-song cross-similarity matrix with convolutional neural network. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 396–400
9. Panagiotou V, Mitianoudis N (2013) PCA summarization for audio song identification using Gaussian mixture models. In: 2013 18th international conference on digital signal processing (DSP). IEEE, pp 1–6
10. Wang Q, Liu G, Guo Z, Guo J, Chen X (2011) Structural fingerprint based hierarchical filtering in song identification. In: 2011 IEEE international conference on multimedia and expo. IEEE, pp 1–4
11. Maddage NC (2006) Automatic structure detection for popular music. Ieee Multimedia 13(1):65–77
12. Dutta S (2015) Geo environmental specificities of Bengali folk song. PhD Thesis
13. Mukherjee H, Dhar A, Obaidullah SM, Santosh KC, Phadikar S, Roy K (2019) Linear predictive coefficients-based feature to identify top-seven spoken languages. Int J Pattern Recogn Artif Intell 2058006
14. Mukherjee H, Obaidullah SM, Phadikar S, Roy K (2018) MISNA-A musical instrument segregation system from noisy audio with LPCC-S features and extreme learning. Multimedia Tools Appl 77(21):27997–28022
15. Mukherjee H, Dhar A, Obaidullah SM, Santosh KC, Phadikar S, Roy K (2018) Identification of top-3 spoken Indian languages: an ensemble learning-based approach. In: 2018 fourth international conference on research in computational intelligence and communication networks (ICRCICN). IEEE, pp 135–140
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11(1):10–18

Shruti Detection Using Machine Learning and Sargam Identification for Instrumental Audio

Muktharunisa Chapparband, Megha Gururaj Kulkarni, D. K. Sameeksha, A. Vijaya Krishna, and Ashwini Bhat

Abstract In the present-day scenario, most of the things are getting digitized. Music being one of the most celebrated art types is learnt by sundry people and taught by several musicians. Thus, this paper aims at helping them with the advanced technology we have today. In order to play a musical instrument, knowing the Shruti and Sargam (terms used to refer to tone and notes, respectively, in Indian Classical Music) of the music being played is very important. Trained musicians can identify the Shruti and Sargam by listening to the music but beginners cannot. Thus, for the beginners to play any instrument, the Shruti and Sargam should be known prior. This paper deals by detecting the Shruti using machine learning approach and identifying the Sargam of the polyphonic audio which will assist the musicians in numerous ways. It also has vast significance in the field of music information retrieval.

Keywords Indian classical music · Music information retrieval · Musical instrument · Shruti · Sargam · Tone · Notes

1 Introduction

Music is one of the extensively enjoyed content types since ages, existing in many varied representations. When it comes to technicality of music, it contains many details such as tone, notes, leading instrument, genre and many more. For a musician, all these terms might sound familiar, but for a layman, it is hard to understand these terminologies. As in this digital epoch, with many of the things being digitized, it is worthwhile to bring the idea of digitization to the music domain, so that music information is available at the user's fingertips. In the music information retrieval (MIR) field, several research problems of interest involve the automatic elucidation

M. Chapparband · M. G. Kulkarni (✉) · D. K. Sameeksha · A. Vijaya Krishna · A. Bhat
PES University, Bangalore, India

A. Vijaya Krishna
e-mail: vijayakrishna@pes.edu

A. Bhat
e-mail: ashwinib@pes.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_13

of properties of music signals, making use of the concepts that are understood by humans. For this, problems that can be solved by automated systems are derived [1].

Shruti and Swara are the most basic elements that constitute any music signal. Shruti, which is also referred to as quarter tones, is the steady periodic sound that the human ear can detect and a singer or musical instrument can produce.

Swara is a term that implies a note in the successive steps of the octave. The collection of Swaras is called Sargam. Each Swara is associated with a frequency. In Western music, there is a one–one mapping of frequency to note. On contrary to this, Indian Classical Music uses a relative mapping of frequency to note. The frequency that corresponds to the Shruti is termed as base frequency. The frequencies corresponding to all Swaras are relative to the base frequency. The relation between base frequency and other frequencies will be shown in the procedure that follows.

In this paper, a machine learning approach is put forth for Shruti detection. The identified Shruti is used for the Sargam identification of music.

2 Preprocessing

The basic step for building any ML model is to have a proper dataset. As dataset plays a major role in training the model, it should be dealt with carefully. For the project carried out, the standard Dunya dataset (Carnatic and Hindustani Instrumental) is used. These datasets are derived from the Comp Music corpora of Indian art music, for which each recording is associated with an MBID. Corpus comprises three data collections, namely audio recordings, music scores and lyrics, as well as related annotations and metadata [2]. Thus, these annotations and metadata can be used to verify the results obtained later.

Segmentation of audio tracks of the dataset into smaller chunks of fixed length is done as a preprocessing task. It will ensure that all the audio files used are of same length which eases the management of the audio database. Also, audio of smaller length helps in better retrieval of the data. The length of each segment is chosen to be of 30 s duration so that variations in audio features can be clearly studied. A chunk with still smaller length would not fetch much data while a larger length would consist of too much data to be processed. Hence, ideally a length of 30 s is chosen.

3 Shruti Detection

Shruti detection is the most important step in the entire process of Sargam identification, as a misinterpretation of Shruti will lead to incorrect identification of all the Swaras. The following section explains the procedure that is adopted in finding the Shruti of an audio track that uses the machine learning concept. The flowchart for the same is shown in Fig. 1.

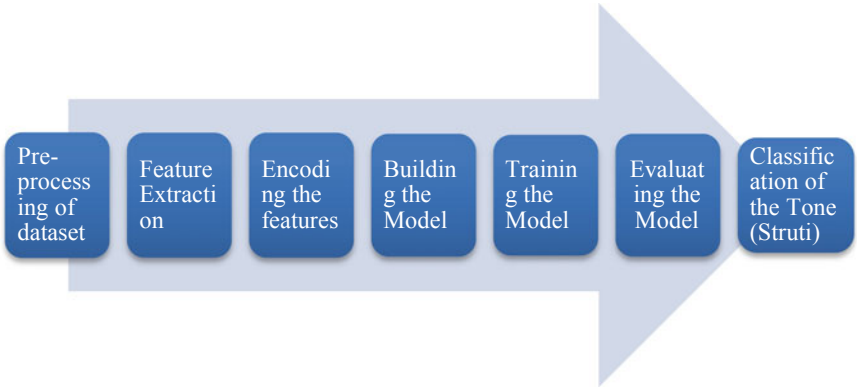


Fig. 1 Flowchart depicting the various steps in Shruti detection using machine learning

In machine learning, the classification is done based on various features of the dataset. So, choosing the appropriate features based on the classification problem and extracting them are of important concern. Feature extraction is a way by which an initial set of raw data is narrowed to feasible groups for processing as it is difficult to deal with raw data directly. The several features under consideration for the Shruti detection are as follows [3, 4].

Envelope-based features include two features which are explained below. The zero crossing rate is the rate of sign changes in a signal. This feature is taken into account in speech recognition and MIR.

Zero crossing rate is calculated by the formula [3]

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_R(s_t s_{t-1} < 0) \tag{1}$$

where s is a signal with duration T, s_t representing the instantaneous value of s at time t, and 1_R is an indicator function. Whenever there is a zero crossing, the signal sign changes and the product $s_t s_{t-1}$ will have a negative value.

RMS of an audio signal is calculated by

$$\sqrt{\frac{1}{N}\sum_{n}|x(n)|^2} \tag{2}$$

Another set of features is spectral-based features which includes three features [3, 5]. A spectral centroid is a quantity that characterizes the spectrum and has a robust relation with brightness of the sound. This is calculated as

$$\text{Spectral Centroid} = \frac{\sum_{k=0}^{N} kF[k]}{\sum_{k=0}^{N} F[k]} \tag{3}$$

where $F[k]$ is the amplitude corresponding to bin k in DFT spectrum.

Spectral bandwidth of order 'p' can be calculated using

$$\left(\sum_k S(k)(f(k) - f_c)^p\right)^{1/p} \tag{4}$$

where $S(k)$ is the spectral magnitude at frequency bin k, $f(k)$ is the frequency at bin k, and f_c is the spectral centroid. When $p = 2$, this is like a weighted standard deviation.

Spectral roll-off is a frequency below which a fixed amount (85%, which is default value in *librosa* library) of the spectral energy is present.

MFCC features are the set of coefficients that constitute a cepstral representation where frequency bands are distributed according to mel scale [3].

Chroma features are used for analyzing the music where the pitch can be determined from the set of twelve pitches.

The next step after extracting the features from the dataset is to train the model. The neural network architecture of the model is depicted in Fig. 2. ReLu is used as an activation function in hidden layers as it avoids vanishing gradients problem, and also it gives more accurate results compared to other activation functions for the dataset considered [6, 7]. The output layer consists of multiple units that correspond

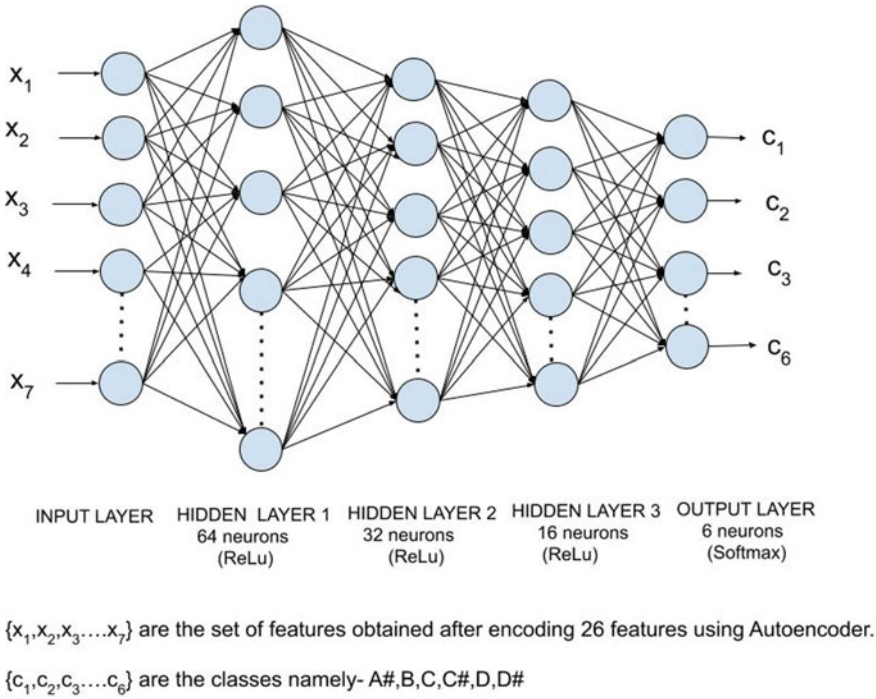


Fig. 2 Neural network architecture of the model

to each Shruti, to be detected. In total six classes, namely A#, B, C, C#, D, D#, are considered in order to make the data samples balanced depending on the number of audio tracks available in the Dunya dataset. As it is a multiclass problem, the softmax function is used as an activation function for the output layer rather than the sigmoid activation function which is applied to binary classification. Specifically, these classes are chosen since most of the audio tracks fall under these ranges of base frequencies. With respect to the learning process of the model, 'Adam' (adaptive moment estimation) optimization algorithm is used. As 'Adam' involves different learning rates set for different parameters while updating the neural weights, the efficiency of this algorithm is high compared to classical stochastic gradient method [8] where learning rate remains constant throughout the process of updating of neural weights. And also, this 'Adam' optimizer achieves good results faster. Cross-entropy is chosen as a loss function, as it is a multiclass problem. This is also called log loss, and this has to be ideally equal to zero to have a perfectly working classification model, which is rare to achieve, instead the loss can be reduced to have a value closer to zero. Once the model is built, the Shruti of the given audio track is detected using this ML model. Confusion matrix and model accuracy are used to analyze the model performance.

4 Sargam Identification

For Sargam identification, knowledge of the base frequency[frequency corresponding to Shadja (Shruti)] is necessary. Once the Shruti for the polyphonic music signal is obtained, it is next used to calculate the base frequency. This involves mapping the Shruti to corresponding frequency using the standard note to frequency mapping table. The frequencies in the standard table are derived using the relation

$$\begin{aligned} \text{frequency} &= 440 \times 2^{n/12} \\ &\text{for n} = -21, -19, \ldots, 27 \end{aligned} \tag{5}$$

4.1 Frequency Identification

Identifying the correct frequencies which correspond to various Swaras present in the signal is very crucial. In this project, this is achieved using two different techniques. In the first method, the concept of sampling, averaging and thresholding is used [9]. The second method is based on the concept of onset detection.

Using the concept of thresholding
This method involves various steps, which are as follows:

Sampling:

The signals used in the real world are analog signals which are continuous in nature. To process these signals in digital machines, they should be converted into digital signals which are discrete in time as well as in amplitude. So, sampling is the process employed to convert a signal from continuous time to discrete time. In this paper, standard sampling frequency of 44.1 kHz is used.

Averaging:

The sampled polyphonic music signal contains a large number of samples as it is sampled at a very high sampling rate. Averaging process is carried out as the next step to smoothen the signal.

Averaging is the process in which the mean value of the amplitude for a set of samples is taken and assigned as one sample. Similarly, the mean value of the next set of samples is considered as a second sample, and so on. If the pace of the music is fast, the notes occur more frequently, and thus an average over 512 samples is taken. On the contrary, if the music is slow, then an average over 1024 samples is taken. The averaged values are used as indicators for the note detection.

Thresholding:

Post-averaging, a process called thresholding, is performed to select the peaks with higher energy levels in the signal, which will finally result in samples that correspond to notes.

Thresholding is done by comparing the averaged amplitudes with a threshold value. Whenever the amplitude value is higher than the threshold, it shows that note has occurred and the corresponding averaged sample is considered. Finally, the corresponding frequency for every note is calculated using autocorrelation. The original sampled signal is used as the input to the autocorrelation function, with two more parameters indicating the start and end index for each frequency calculation. Thus, all the frequencies corresponding to the notes present in the given audio signal are found.

In this method, every time depending on the pace of the music and intensity of the musical notes, threshold value and the number of samples for averaging need to be set dynamically to make sure that no note is missed. This drawback is overcome by using the music onset detection method.

Using Onset Detection

Music onset detection plays an indispensable role in music signal processing and has widespread applications in the field of MIR. Other applications include automatic music transcription, beat-tracking and many more [10].

The detection of a moment when a discrete event begins in a musical signal is termed as onset detection [10]. Onset detection locates the local maxima (i.e., the time when the energy goes high) by picking the peaks in an onset strength envelope. These peaks correspond to the notes. Onset detection is a process associated with the time domain. Thus, using this concept, all the peaks present in a music signal

are detected and the timing (in seconds) at which these peaks occur is retrieved. To convert the timing information into discrete time domain, the sample index of these timestamps is calculated. Later these samples are correlated and used to determine the frequencies.

Onset detection method is a better way of identifying the occurrence of peaks that correspond to the notes than the former method since in this method the number of samples for averaging need not be decided priorly depending on the pace of the music. Also, the threshold value is not required to be set. Onset detection will detect all the peaks of a given signal, ensuring no notes are missed, and finally the corresponding frequencies are obtained.

4.2 Frequency Mapping to Swara

After Shruti detection and frequency identification, mapping these frequencies to corresponding Swaras is the final step. As mentioned earlier, the Indian Musical System uses a relative mapping of the Swaras. Hence, irrespective of the octaves, the frequencies associated with each Swara are related to the base frequency by specific ratios [11, 12].

The Shruti found earlier using machine learning approach is used to determine the base frequency. The Swara corresponding to the base frequency is taken to be the 'Sa' note. Using Table 1, the frequencies for all the other Swaras are calculated and compared with the frequencies identified in the previous step, and eventually, these frequencies are mapped onto the corresponding Swaras to obtain the Sargam of the polyphonic music signal.

Table 1 Relative value of the frequency of each Swara with respect to the frequency of the Shruti (base frequency)

S. No.	Swara (note)	Relative value	Fraction
1	Sa (Shruti)	1	1
2	Ri1	256/243	1.053
3	Ri2	9/8	1.1250
4	Ga1	32/27	1.1851
5	Ga2	81/64	1.2656
6	Ma1	4/3	1.3333
7	Ma2	729/512	1.4238
8	Pa	3/2	1.5
9	Dha1	128/81	1.5802
10	Dha2	27/16	1.6875
11	Ni1	16/9	1.7777
12	Ni2	243/128	1.8984

5 Results

The results are obtained for the Hindustani and Carnatic Instrumental tracks from the Dunya dataset.

After building the machine learning model, the confusion matrix is plotted for the evaluation of the accuracy of the predicted values. A total of 1200 audio samples are taken as part of the training and testing data. Twenty percentage of this is considered for testing. The confusion matrix is depicted in Fig. 3.

The results are obtained for several input files from the Dunya dataset. The results for two files, one from Hindustani instrumental and the other from Carnatic instrumental, are demonstrated in the paper. The audio files considered are audio-1: "Hariprasad Chaurasia - Dhun Sindh Bhairavin" and audio-2: "Kadri Gopalnath Gnana Vinayakane". The results are obtained for the first segment which is the initial 30 s of both the audios. All the results corresponding to these audios are shown in Figs. 4, 5, 6, 7, 8, 9, 10, and 11.

The determination of peaks using the onset detection method for the audio-1 is depicted in Fig. 4.

The Shruti, frequencies and Sargam detected for the audio-1 are depicted in Figs. 5, 6 and 7.

The determination of peaks using the onset detection method for the audio-2 is depicted in Fig. 8.

The Shruti, frequencies and Sargam detected for the audio-2 are depicted in Figs. 9, 10 and 11.

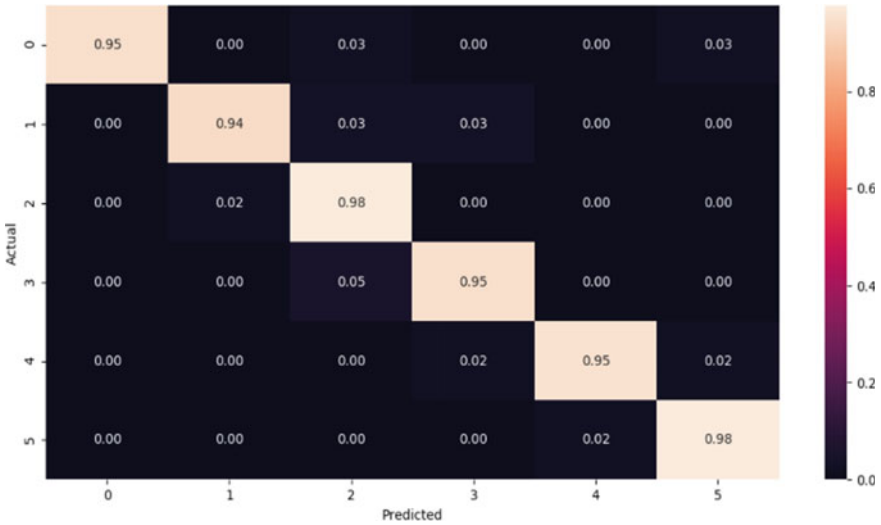


Fig. 3 Confusion matrix of the model with six Shrutis as classes

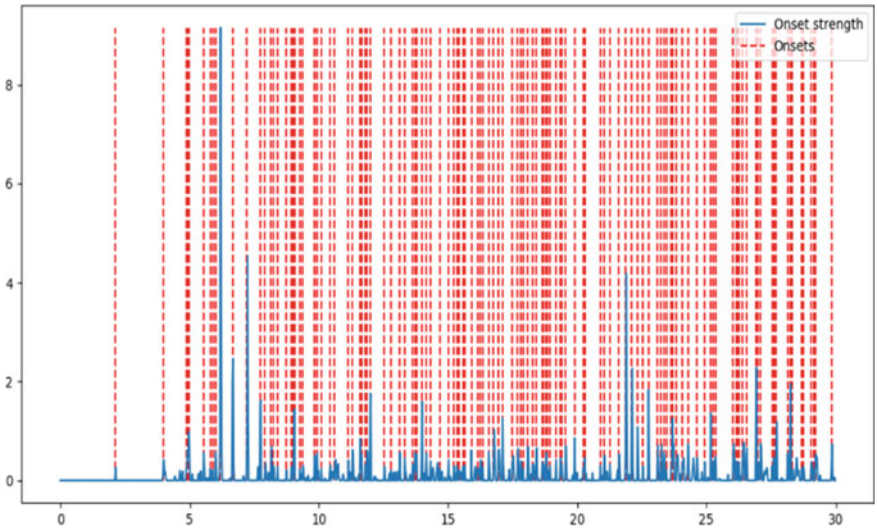


Fig. 4 Peak picking in onset detection for audio-1

```
Shruti for the given track is: D


The Frequencies are:
 [735.0, 142.72, 227.32, 760.34, 227.32, 143.18, 143.18, 143.65, 143.18, 216.18,
 144.12, 107.82, 290.13, 143.65, 383.48, 531.33, 565.38, 572.73, 572.73, 572.73,
 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73,
 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73, 572.73,
 572.73, 572.73, 572.73, 537.8, 572.73, 531.33, 428.16, 347.24, 347.24, 324.26,
326.67, 324.26, 339.23, 339.23, 341.86, 347.24, 344.53, 344.53, 347.24, 339.23,
339.23, 341.86, 336.64, 344.53, 347.24, 347.24, 339.23, 344.53, 341.86, 339.23,
344.53, 339.23, 344.53, 344.53, 344.53, 341.86, 324.26, 344.53, 339.23, 143.65,
143.18, 143.18, 141.8, 228.5, 152.07, 144.12, 143.65, 143.65, 143.65, 142.72, 28
6.36, 286.36, 580.26, 580.26, 580.26, 580.26, 292.05, 294.0, 304.14, 334.09, 341
.86, 341.86, 347.24, 341.86, 341.86, 341.86, 355.65, 376.92, 329.1, 339.23, 358.
54, 432.35, 143.18, 383.48, 142.26, 355.65, 339.23, 380.17, 383.48, 143.18, 326.
67, 319.57, 288.24]
```

Fig. 5 Output showing the Shruti and various frequencies for audio-1

```
The Sargam notes are:
 ['Sa', 'Sa', 'Dha1', 'Sa', 'Dha1', 'Sa', 'Sa', 'Sa', 'Sa', 'Pa', 'Sa', 'Sa', 'S
a', 'Sa', 'Ma1', 'Ni1', 'Ni1', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', '
Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa
', 'Sa', 'Sa', 'Sa', 'Sa', 'Ni1', 'Sa', 'Ni1', 'Pa', 'Ga1', 'Ga1', 'Ri2', 'Ri2',
 'Ri2', 'Ri2', 'Ri2', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ri2', 'Ri2', 'Ga1', 'R
i2', 'Ga1', 'Ga1', 'Ga1', 'Ri2', 'Ga1', 'Ga1', 'Ri2', 'Ga1', 'Ri2', 'Ga1', 'Ga1'
, 'Ga1', 'Ga1', 'Ri2', 'Ga1', 'Ri2', 'Sa', 'Sa', 'Sa', 'Sa', 'Dha1', 'Ri2', 'Sa'
, 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Sa', 'Ri2',
 'Ri2', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ga1', 'Ri2', 'Ri2', 'G
a1', 'Pa', 'Sa', 'Ma1', 'Sa', 'Ga1', 'Ri2', 'Ga1', 'Ma1', 'Sa', 'Ri2', 'Ri2', 'S
a']
```

Fig. 6 Output showing Sargam notes for audio-1

```
['Sa', '__', 'Dha1', 'Sa', 'Dha1', 'Sa', '__', '__', '__', 'Pa', 'Sa', '__', '__
', '__', 'Ma1', 'Ni1', '__', 'Sa', '__', '__', '__', '__', '__', '__', '__', '__
', '__', '__', '__', '__', '__', '__', '__', '__', '__', '__', '__', '__', '__', '__
', '__', '__', '__', '__', '__', 'Ni1', 'Sa', 'Ni1', '__', 'Pa', 'Ga1', '__', 'Ri2', '__', '__
', '__', '__', 'Ga1', '__', '__', '__', '__', '__', 'Ri2', '__', 'Ga1', 'Ri2', 'Ga1',
'__', '__', 'Ri2', 'Ga1', '__', 'Ri2', 'Ga1', 'Ri2', 'Ga1', '__', '__', '__', 'R
i2', 'Ga1', 'Ri2', 'Sa', '__', '__', '__', 'Dha1', 'Ri2', 'Sa', '__', '__', '__'
, '__', '__', '__', '__', '__', '__', '__', '__', '__', '__', 'Ri2', '__', 'Ga1', '__'
, '__', '__', '__', '__', '__', '__', 'Ri2', '__', 'Ga1', 'Pa', 'Sa', 'Ma1', 'Sa
', 'Ga1', 'Ri2', 'Ga1', 'Ma1', 'Sa', 'Ri2', '__', 'Sa']
```

Fig. 7 Music script after eliminating the redundant Swaras for audio-1

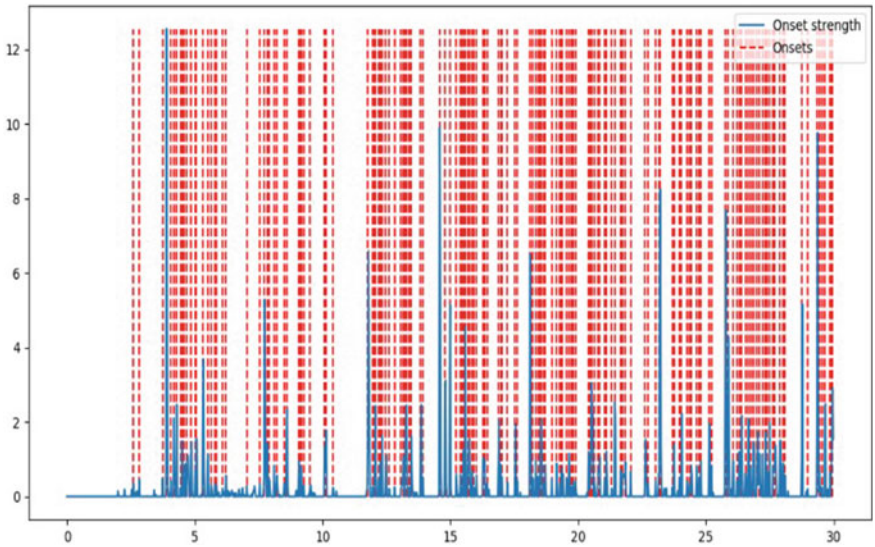


Fig. 8 Peak picking in onset detection for audio-2

```
Shruti for the given track is: A#


The Frequencies are:
 [367.5, 364.46, 352.8, 308.39, 352.8, 469.15, 537.8, 454.64, 479.35, 352.8, 308
.39, 297.97, 370.59, 306.25, 297.97, 295.97, 295.97, 295.97, 295.97, 295.97, 295
.97, 295.97, 297.97, 286.36, 223.86, 232.11, 223.86, 235.83, 119.51, 213.04, 358
.54, 286.36, 112.21, 72.18, 221.61, 246.37, 295.97, 238.38, 279.11, 226.15, 228.
5, 173.62, 230.89, 156.38, 163.94, 232.11, 233.33, 290.13, 169.62, 87.67, 157.5,
 156.94, 152.6, 178.54, 237.1, 239.67, 294.0, 308.39, 352.8, 551.25, 450.0, 474.
19, 350.0, 352.8, 306.25, 275.62, 295.97, 350.0, 304.14, 295.97, 295.97, 279.11,
 297.97, 300.0, 297.97, 310.56, 297.97, 297.97, 223.86, 272.22, 223.86, 180.0, 1
73.62, 158.06, 160.95, 175.0, 223.86, 234.57, 242.31, 233.33, 232.11, 238.38, 25
7.89, 250.57, 235.83, 117.6, 237.1, 237.1, 237.1, 237.1, 222.73, 280.89, 259.41,
 234.57, 226.15, 175.0, 158.63, 177.82, 222.73, 295.97, 235.83, 256.4, 317.27, 3
19.57, 217.24, 234.57, 282.69, 234.57, 175.0, 158.63, 176.4, 225.0, 290.13, 310.
56, 352.8, 249.15, 308.39, 225.0, 233.33, 254.91, 290.13, 238.38, 279.11, 221.61
, 175.7, 222.73, 156.38, 190.09, 240.98, 279.11, 352.8, 441.0, 558.23, 454.64, 1
72.27, 288.24, 370.59, 300.0, 259.41, 312.77, 292.05, 265.66, 347.24, 469.15, 45
9.38, 459.38, 350.0]
```

Fig. 9 Output showing the Shruti and various frequencies for audio-2

```
The Sargam notes are:
 ['Dha1', 'Dha1', 'Pa', 'Ma1', 'Pa', 'Sa', 'Sa', 'Sa', 'Sa', 'Pa', 'Ma1', 'Ga2',
 'Dha1', 'Ma1', 'Ga2', 'Ga2', 'Ga2', 'Ga2', 'Ga2', 'Ga2', 'Ga2', 'Ga2', 'Ga2', '
Ga2', 'Ni2', 'Sa', 'Ni2', 'Sa', 'Sa', 'Ni2', 'Pa', 'Ga2', 'Sa', 'Sa', 'Ni2', 'Ri
2', 'Ga2', 'Sa', 'Ga2', 'Ni2', 'Sa', 'Pa', 'Sa', 'Ma1', 'Ma1', 'Sa', 'Sa', 'Ga2'
, 'Ma1', 'Sa', 'Ma1', 'Ma1', 'Ma1', 'Pa', 'Sa', 'Sa', 'Ga2', 'Ma1', 'Pa', 'Sa',
'Ni2', 'Sa', 'Pa', 'Pa', 'Ma1', 'Ga2', 'Ga2', 'Pa', 'Ma1', 'Ga2', 'Ga2', 'Ga2',
'Ga2', 'Ga2', 'Ga2', 'Ma1', 'Ga2', 'Ga2', 'Ni2', 'Ga2', 'Ni2', 'Dha1', 'Pa', 'Ma
1', 'Ma1', 'Pa', 'Ni2', 'Sa', 'Ri2', 'Sa', 'Sa', 'Sa', 'Ri2', 'Ri2', 'Sa', 'Sa',
 'Sa', 'Sa', 'Sa', 'Sa', 'Ni2', 'Ga2', 'Ri2', 'Sa', 'Ni2', 'Pa', 'Ma1', 'Pa', 'N
i2', 'Ga2', 'Sa', 'Ri2', 'Ma1', 'Ma1', 'Ni2', 'Sa', 'Ga2', 'Sa', 'Pa', 'Ma1', 'P
a', 'Ni2', 'Ga2', 'Ma1', 'Pa', 'Ri2', 'Ma1', 'Ni2', 'Sa', 'Ri2', 'Ga2', 'Sa', 'G
a2', 'Ni2', 'Pa', 'Ni2', 'Ma1', 'Dha1', 'Ri2', 'Ga2', 'Pa', 'Ni2', 'Sa', 'Sa', '
Pa', 'Ga2', 'Dha1', 'Ga2', 'Ri2', 'Ma1', 'Ga2', 'Ri2', 'Pa', 'Sa', 'Sa', 'Sa', '
Pa']
```

Fig. 10 Output showing Sargam notes for audio-2

```
['Dha1', '__', 'Pa', 'Ma1', 'Pa', 'Sa', '__', '__', '__', 'Pa', 'Ma1', 'Ga2', 'D
ha1', 'Ma1', 'Ga2', '__', '__', '__', '__', '__', '__', '__', '__', '__', '__', 'Ni2',
 'Sa', 'Ni2', 'Sa', '__', 'Ni2', 'Pa', 'Ga2', 'Sa', '__', 'Ni2', 'Ri2', 'Ga2', '
Sa', 'Ga2', 'Ni2', 'Sa', 'Pa', 'Sa', 'Ma1', '__', 'Sa', '__', 'Ga2', 'Ma1', 'Sa'
, 'Ma1', '__', '__', 'Pa', 'Sa', '__', 'Ga2', 'Ma1', 'Pa', 'Sa', 'Ni2', 'Sa', 'P
a', '__', 'Ma1', 'Ga2', '__', 'Pa', 'Ma1', 'Ga2', '__', '__', '__', '__', '__',
'Ma1', 'Ga2', '__', 'Ni2', 'Ga2', 'Ni2', 'Dha1', 'Pa', 'Ma1', '__', 'Pa', 'Ni2',
 'Sa', 'Ri2', 'Sa', '__', '__', 'Ri2', '__', 'Sa', '__', '__', '__', '__', '__',
 'Ni2', 'Ga2', 'Ri2', 'Sa', 'Ni2', 'Pa', 'Ma1', 'Pa', 'Ni2', 'Ga2', 'Sa', 'Ri2',
 'Ma1', '__', 'Ni2', 'Sa', 'Ga2', 'Sa', 'Pa', 'Ma1', 'Pa', 'Ni2', 'Ga2', 'Ma1',
'Pa', 'Ri2', 'Ma1', 'Ni2', 'Sa', 'Ri2', 'Ga2', 'Sa', 'Ga2', 'Ni2', 'Pa', 'Ni2',
'Ma1', 'Dha1', 'Ri2', 'Ga2', 'Pa', 'Ni2', 'Sa', '__', 'Pa', 'Ga2', 'Dha1', 'Ga2'
, 'Ri2', 'Ma1', 'Ga2', 'Ri2', 'Pa', 'Sa', '  ', '  ', 'Pa']
```

Fig. 11 Music script after eliminating the redundant Swaras for audio-2

Thus, during the process, it was observed that the frequencies identified with the onset detection method were more accurate compared to the method of thresholding. The reason being, averaging over 512 samples resulted in loss of information when the notes occurred at higher pace. At the same time, reducing the number of samples leads to overhead in the computation. Hence, the onset detection method was used to obtain the above results.

6 Conclusions

The Shruti for the given audio track is detected using the machine learning model successfully. The confusion matrix plotted for the model shows a high degree of accuracy for the dataset chosen. Detected Shruti is mapped onto the associated frequency. This gives the base frequency. The various frequencies present in the music signal are identified using two different techniques. Finally, these frequencies are then mapped onto the corresponding Swaras. The results obtained are verified against the annotations and metadata present with the Dunya dataset.

References

1. Kim J, Urbano J, Liem CCS et al (2020) One deep music representation to rule them all? A comparative analysis of different representation learning strategies. Neural Comput Appl 32:1067–1093. https://doi.org/10.1007/s00521-019-04076-1
2. Porter A, Sordo M, Serra X (2013) Dunya: a system for browsing audio music collections exploiting cultural context. In: Proceedings of the international society for music information retrieval conference, pp 101–106
3. Chandan SV, Naik MR, Ashwini, Krishna AV (2019) Indian instrument identification from polyphonic audio using KNN classifier. In: 2019 international conference on wireless communications signal processing and networking (WiSPNET), Chennai, India
4. Karunakaran N, Arya A (2018) A scalable hybrid classifier for music genre classification using machine learning concepts and spark. In: 2018 international conference on intelligent autonomous systems (ICoIAS), Singapore
5. Kini S, Gulati S, Rao P (2011) Automatic genre classification of North Indian devotional music. In: 2011 national conference on communications (NCC), Bangalore
6. Stursa, Dolezel P (2019) Comparison of ReLU and linear saturated activation functions in neural network for universal approximation. In: 2019 22nd international conference on process control (PC19), Strbske Pleso, Slovakia
7. Shen G, Yuan Y (2019) On theoretical analysis of single hidden layer feedforward neural networks with Relu activations. In: 2019 34th youth academic annual conference of Chinese Association of Automation (YAC), Jinzhou, China
8. Bock S, Weiß M (2019) A proof of local convergence for the adam optimizer. In: 2019 international joint conference on neural networks (IJCNN), Budapest, Hungary
9. Patel J, Gopi ES (2015) Musical notes identification using digital signal processing. Procedia Comput Sci 57:876–884. https://doi.org/10.1016/j.procs.2015.07.499
10. Zhou R, Reiss JD (2011) Music onset detection. In: Wang W (ed) Machine audition: principles, algorithms and systems. IGI Global, pp 297–316. https://doi.org/10.4018/978-1-61520-919-4.ch012
11. Sinith MS, Tripathi S, Murthy KVV (2015) Real-time Swara recognition system in Indian Music using TMS320C6713. In: 2015 international conference on advances in computing, communications and informatics (ICACCI), Kochi
12. Akant K (2019) Measuring frequencies of Shrutis in Indian Classical Music. In: 2019 9th international conference on emerging trends in engineering and technology—signal and information processing (ICETET-SIP-19), Nagpur, India

Addressing the Recitative Problem in Real-Time Opera Tracking

Charles Brazier and Gerhard Widmer

Abstract Robust real-time opera tracking (score following) would be extremely useful for many processes surrounding live opera staging and streaming, including automatic lyrics displays, camera control, or live video cutting. Recent work has shown that, with some appropriate measures to account for common problems such as breaks and interruptions, spontaneous applause, various noises, and interludes, current audio-to-audio alignment algorithms can be made to follow an entire opera from beginning to end, in a relatively robust way. However, they remain inaccurate when the textual content becomes prominent against the melody or music—notably, during *recitativo* passages. In this paper, we address this specific problem by proposing to use two specialized trackers in parallel, one focusing on music- and the other on speech-sensitive features. We first carry out a systematic study on speech-related features, targeting the precise alignment of corresponding recitatives from different performances of the same opera. Then, we propose different solutions, based on pre-trained music and speech classifiers, to combine the two trackers in order to improve the global accuracy over the course of the entire opera.

1 Introduction

Score following aims at aligning a music performance to its corresponding score in real time, in order to retrieve the score position currently played by the musician(s) during a live performance. Since its start in the 1980s [1], score following research

C. Brazier (✉)
Institute of Computational Perception, Johannes Kepler University,
Altenberger Straße 69, 4040 Linz, Austria
e-mail: charles.brazier@jku.at

G. Widmer
Institute of Computational Perception and LIT AI Lab, Linz Institute of Technology,
Altenberger Straße 69, 4040 Linz, Austria
e-mail: gerhard.widmer@jku.at

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_14

has produced increasingly sophisticated music trackers, capable of reliably following complex solo performances and even full orchestras [2], supporting such applications as automatic page turning for pianists [3], live performance visualization [4], or score viewing and automatic contextualization in orchestra concerts [5, 6].

Reliably tracking complex musical stage works such as full operas, however, is an even more difficult task. Operas are highly complex and heterogeneous works, involving not only a full orchestra superimposed with a complex mixture of singing voices (of highly expressive and varying singing or speaking style, generally with strong vibrato), but also various acting breaks and interruptions, spontaneous applause, stage noises, interludes, etc. At the same time, robust live opera tracking systems would be extremely useful for many processes surrounding live opera staging and streaming, including automatic lyrics displays, camera control, or live video cutting.

A first step toward robust tracking of full operas was presented in [7]. The approach to score following taken there makes use of an audio-to-audio alignment strategy and aligns a live opera, the *target*, with a *reference recording*, which was aligned to the printed score before and serves as a proxy to the score. This method has the advantage of aligning two audio signals that contain realistic sounds and that can be expected to be more similar, compared to using an artificial reference audio computed from a symbolic score (given as MIDI or Music XML). It also obviates the need for full symbolic opera scores, which are generally not available. However, this method does require manual annotation beforehand. In [7], the real-time alignment between live target and reference is then computed with an adapted version of the *Online Dynamic Time Warping (OLTW)* algorithm [8], combined with applause, speech, and music classifiers that exert top-down control on the tracker in specific situations.

However, some open problems remain, notably tracking *recitativo* sections, which are often introduced between songs to progress fast in the story. Recitatives are strongly acted and are dominated by the singers' voices. The singers are relatively free to impose their own rhythm to best match their acting on stage. The movements on stage directly alter the audio quality that is recorded by fixed microphones. Also, there may be high variability in vocal style, with singers singing, speaking, whispering, or laughing. In addition, the accompaniment is usually underspecified in the score and filled in or improvised in different ways (and even on different instruments) in different recordings.

In this paper, after presenting the real-world musical material that we will be working with (Sect. 2), we will present empirical results demonstrating the difficulty of recitative tracking (Sect. 3), carry out a systematic search for audio features that are more effective for tracking recitatives (Sect. 4), and investigate several strategies for combining music- and speech-sensitive based trackers, using pretrained audio classifiers (Sect. 5). The investigations are based on several full-length recordings of W.A. Mozart's opera *Don Giovanni*, recorded live in an opera hall.

2 Data Description

As opera research datasets including recordings and corresponding annotations are not yet freely available, we had to compile our own dataset. As our first object of study, we chose the opera *Don Giovanni* (considered by some to be the "opera of all operas"). We collected and manually annotated three complete recordings of this work, one from a commercial CD (to serve as the proxy to the score, the *reference performance*), the other two recent live recordings from the opera hall, provided to us by the *Vienna State Opera*.[1] The latter two will be the target performances that we wish to align/track online. There are a few pieces (arias, etc.) that are present in the reference but not in the target performances; these will be ignored for this study. The two target performances have the same structure.

More specifically (see also Table 1), the reference is a CD recording made by Herbert von Karajan in 1985. To link this performance to the score, we manually annotated the recording at the bar level, involving 5304 bar annotations (2866 for the first act, 2438 for the second). The two tracking targets are recordings of live performances from the Vienna State Opera in 2018 and 2019, conducted by Ádam Fischer and Antonello Manacorda, respectively, and performed by entirely different casts of singers. Again, for evaluation purposes, these had to be manually annotated, resulting in 10,608 new bar annotations.

For the study of this paper, we isolate from each recording the passages that contain only *recitativo* sections. This subset represents between 16 and 18% of the total duration of each recording. In the experiments, the Fischer live recording will be used for the feature optimisation study; the Manacorda will then serve as the independent test set.

In total, the dataset comprises more than 9 h of real-world opera music and involves a significant number of 15,912 manual bar-level annotations linking audio and pdf score. The score of *Don Giovanni* comprises some 530 pages, in the Bärenreiter Urtext version (also available online, thanks to the *Salzburg Mozarteum Foundation*[2]).

3 Baseline Tracking Evaluation

The opera tracker described in [7] combines an Online Time Warping (OLTW) algorithm with audio event detectors (classifiers detecting the presence of applause, music, and speech) that can halt the tracking process during transitions between parts. This makes the tracker quite robust to spontaneous events, such as applause, breaks, and interludes, which frequently occur at transitions between arias or recitatives. As input features to the Dynamic Time Warping (DTW) alignment algorithm, and based on a systematic study on classical orchestra recordings [9], 120 Mel-Frequency

[1] https://www.wiener-staatsoper.at/en/.

[2] https://dme.mozarteum.at/DME/nma/.

Table 1 Dataset used in the audio-to-audio alignment strategy

Conductor	Orchestra	Year	Duration	Dur. Recitatives	Role
H.v. Karajan	Berliner Philharmoniker	1985	2:57:53	0:30:03	Reference
Á. Fischer	Orchester der Wiener Staatsoper	2018	3:12:54	0:34:58	Target
A. Manacorda	Orchester der Wiener Staatsoper	2019	3:07:09	0:30:40	Target

Table 2 SoTA tracker: tracking error on full-length Don Giovanni, on the music subset only and on the recitative subset only

Recording	Dataset	Mean (s)	≤1 s(%)	≤2 s(%)	≤5 s(%)
Fischer	Full	0.8	91.8	95.0	97.3
	Music	0.2	97.2	99.2	99.9
	Recitative	1.9	66.1	75.3	88.9
Manacorda	Full	0.6	90.1	94.5	97.9
	Music	0.2	96.2	98.7	99.7
	Recitative	1.5	62.0	77.2	92.9

Absolute durations in Fischer: full: 03:12:54; music: 02:37:56; recitative: 00:34:58. Manacorda: full: 03:07:09; music: 02:36:29; recitative: 00:30:40

Cepstral Coefficients (MFCC) features are extracted from both recordings; the first 20 are discarded, giving a feature vector of length 100 at each audio frame. Fast Fourier Transform (FFT) parameters are a window size of 20 ms and a hop size of 10 ms.

In order to highlight difficulties related to recitative tracking, we separately evaluate three different alignment scenarios: (1) alignment on the full-length target performances; (2) alignment on the music excluding all *recitativo* sections (which leaves 4665 annotated bars of music); and (3) alignment only on *recitativo* sections (639 bars).

Table 2 reports the tracking error on the Fischer and Manacorda performances, when aligned with the Karajan as the reference. The evaluation metrics used are standard in the field of score following [10]: for each alignment, we report the mean tracking error in seconds over all bar annotations, as well as the cumulative proportion of bar positions tracked with an error of less than 1 s, 2 s, and 5 s. We observe the same effect on both performances. For example, in the Fischer alignment, we have a mean tracking error of 0.8 s, with 91.8% of the bars detected with an error below 1 s, 95.0% below 2 s, and 97.3% below 5 s. Tracking on the music-only part gives the lowest mean error and the highest percentages, implying a more precise alignment. The results on the recitative-only part are significantly worse: the number of bars

detected with error below a second decreases to 66.1 and 62.0% for the Fischer and Manacorda, respectively.

Our goal in the following is to find a set of features that are better suited to accurate recitative alignment/tracking. In Sect. 5, we will investigate strategies for combining music- and recitative-focused features and trackers.

4 Specific Features for the Recitative Problem

Our first step toward improving recitative tracking is to search for a set of features that are better suited to modeling passages with predominant voice content. Operatic singing voices are usually clustered into six categories (bass, baritone, tenor, alto, mezzo-soprano, and soprano), depending on their f0 (pitch) ranges. The frequency ranges extend from 73 to 1500 Hz [11, 12]. The features used in our baseline tracker and described in [7] cover the frequency range [659; 22050 Hz]. Even if they enable accurate tracking in orchestra performances [9], they may not be appropriate to track recitatives.

Aligning two performances include the same linguistic content, but whose vocal styles differ is not trivial. Apart from vocal effects such as vibratos or tremolos, a noticeable difference between speech and singing characteristics is the presence of a peak in the singing spectrum that is situated around 3 kHz, namely the 'singer's formant' [13]. Thus, and in order to get features robust to vocal style changes, one option is to restrict the frequency axis to a limit of 3 kHz to avoid this phenomenon and to focus on the frequency range mentioned previously.

4.1 Feature Search Space

In our search for an appropriate set of features, we again turn to MFCCs and start a systematic search for a parametric setup optimized for recitative tracking. MFCCs are pervasively used in speech recognition and audio-to-lyrics alignment, and are already the features used in the baseline tracker. We extract MFCC features from three types of spectra that can be calculated in real time. The first spectrum is the standard one (called *Spec* in the following), with a window size of 20 ms and a hop size of 10 ms. The two others are two different methods that estimate the *spectral envelope* of the spectrum, the latter representing the phoneme information in a Source/Filter model (see equation (7) in [14]). One method computes the Linear Predictive Coefficients (LPC), a robust feature that has already proven successful in the speech-to-singing alignment domain [15]. The other computes the True Envelope (TE), which estimates the spectral envelope in an iterative way, and has been used in a singing-voice-to-score alignment task [14]. From these two different spectral envelopes, we then extract the corresponding MFCCs.

To find the best set of parameters for the MFCC features, a grid search is conducted over the following parameter space:

- Sampling Rate: According to the previous considerations, we propose downsampling the initial audio signal. Six different **#sr** values are tested: 1500, 3000, 6000, 12,000, 24,000, and 44,100 Hz.
- Number of MFCCs: This number changes the resolution of the features. Six **#MFCC** values are proposed: 25, 50, 75, 100, 150, and 200.
- Discarding Factor: This refers to the number of first (low-order) MFCCs to be discarded from the set. The factor was under study in [9]. Two **#skip** values are proposed: 0 and 5.

The evaluation criterion for the grid search is the alignment error on the recitatives in the Fischer recording. The Manacorda is not used in this process, so that it can serve as an independent test set later.

In total, 216 experiments have been run, involving three types of features, six values of sampling rate, two discarding factors, and six numbers of MFCC.

4.2 Optimization on Fischer Recitatives

Table 3 shows the alignment errors on the recitative subset of the Fischer performance (639 bars), for the 12 best feature parameter settings. As a reference, the *baseline* features are the features used in our baseline tracker [7].

Table 3 Feature parameter grid search on the recitative subset of *Don Giovanni* conducted by Ádam Fischer

Feature	#sr	#MFCC	#skip	Mean (ms)	≤1 s (%)
Baseline	44,100	100	20	1915	66.1
LPC	1500	25	0	**955**	**76.5**
TE	1500	75	0	965	76.4
TE	1500	100	0	965	76.4
TE	1500	150	0	965	76.4
Spec	1500	75	0	966	76.4
TE	1500	200	0	976	76.2
Spec	1500	100	0	977	76.2
Spec	1500	150	0	977	76.2
Spec	1500	200	0	977	76.2
Spec	1500	50	0	979	76.2
Spec	1500	25	0	989	76.2
LPC	1500	50	0	958	76.1

Only the best results (percentage of bars detected with error ≤1 s higher than 76%) are shown

Table 4 Validating the features on the recitative subset of the Manacorda performance

Feature	#sr	#MFCC	#skip	Mean (ms)	≤1 s (%)
Baseline	44,100	100	20	1503	62.0
LPC	1500	25	0	**1023**	**69.7**

The significance of bold in Tables 3, 4, 5, and 6 indicates the best score in the two criteria. For the mean, the lower the better. For the mean ≤1 s, the higher the better

First, we note that all of these top 12 feature settings give relatively close results, in terms of mean alignment error and alignment precision. All features clearly outperform the baseline, reducing the mean error by one second and improving the percentage of precisely detected bars by at least 10% points.

All the best results involve the same low sampling rate 1500 Hz. This is in accordance with the argument made in the previous section. It allows not to include the 'singing formant' phenomenon in the different spectra and to zoom in on the frequency range corresponding to the opera singers.

Finally, we note that increasing the discarding factor to 5 significantly decreases the alignment performance (not shown in the table), meaning that the lower order coefficients are important for recitative tracking. This relates to the point that the first MFCC coefficients are representative to the spectral shape [16], which contains most of the phoneme information, according to the source/filter model.

A final observation (not included in Table 3) is that the *maximum alignment error* over all these recitatives dropped from 26.8 s in the baseline case to 13.1 s when using the optimized features. Thus, the new features may help reduce the risk of the tracker getting completely lost in difficult passages.

4.3 Validating on Manacorda Recitatives

Based on Table 3, we conclude that the best performances were obtained in extracting a smoothed spectral envelope based on LPC with a sampling rate 1500 Hz, 25 MFCC features, and a discarding factor of 0. We now validate this optimised feature set on the independent Manacorda recitative set and obtain the results as given in Table 4. The improvement over the baseline is obvious. However, the amount of improvement is somewhat less spectacular than in the Fischer case; also the *maximum* alignment error in this case remains unchanged at 15.0 s, due to one particularly difficult situation in the recitative *Ah ah ah ah, questa è buona*, which includes an improvised interlude of more than 13 s at the beginning of the third bar.

5 Combining Trackers

The goal of this work is to improve the accuracy of a full opera tracker. Based on the results from the previous section, we now have at our disposal two specific sets of features: one that is effective for tracking orchestral music (the MFCCs from our baseline tracker), another that is effective for tracking recitatives. We now present and evaluate different strategies for combining both features in order to improve the overall opera tracking. To this end, we presuppose that we have two reasonably reliable (online) audio classifiers that detect the presence of music and speech, respectively, in an audio stream. Such classifiers, based on yet other sets of specialized features, are already components of our baseline tracker; they were trained on separate annotated training data and are briefly described in [7].

5.1 Combination Strategies

The combined tracking model is illustrated in Fig. 1. From a performance, we extract a set of music-sensitive features and speech-sensitive features. From each type of feature, we compute the alignment between reference and performance, running two specialized OLTW trackers in parallel. At each audio frame, each DTW tracker outputs a cumulative distance vector whose minimal value is the hypothetical current score position. In the following, we propose four simple solutions for combining these. We will use γ_m to refer to the distance vector output by the music tracker and γ_s to the vector of the tracker using the speech-related (recitative) features.

1. Proposed by [14], the *late fusion* strategy calculates the average of the two distance vectors γ_m and γ_s. This strategy has proven to be successful in a hidden semi-

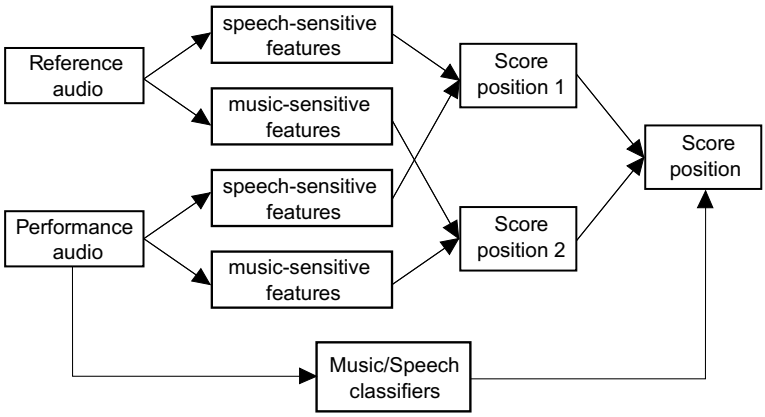


Fig. 1 Integrated model

Markov model. It has the advantage to compensate a weak tracker with the other one:

$$\gamma_{\text{LateFus}} = \frac{\gamma_m + \gamma_s}{2}$$

2. The alternative combination variants that we propose use our trained music/speech audio classifiers. In the *speech-based fusion* strategy, considering that the tracker that aligns speech-sensitive features tends to be more accurate during the presence of speech, we weigh the cumulative vectors by the probability of speech being present in the current signal, derived from the recent predictions of the speech detector:

$$\gamma_{\text{SpeechFus}} = (1 - w_s) \cdot \gamma_m + w_s \cdot \gamma_s$$

   We calculate w_s in two different ways: as the average of the last 25 speech predictions (probabilities) output by the speech detector (which, at a frame rate of 20 ms, gives a temporal context of 500 ms), and as a weighted average with a linear weighting function from 0 to 1, with a context of 1 s.
3. Analogously, the *music-based fusion* strategy mixes in the music-sensitive tracker when music is detected:

$$\gamma_{\text{MusicFus}} = w_m \cdot \gamma_m + (1 - w_m) \cdot \gamma_s$$

   We also compute w_m in two ways.
4. The *music and speech-based* fusion combines the two previous propositions and applies the equation:

$$\begin{aligned}\gamma_{\text{M\&SFus}} &= 0.5 \cdot (w_m + (1 - w_s)) \cdot \gamma_m \\ &\quad + 0.5 \cdot ((1 - w_m) + w_s) \cdot \gamma_s\end{aligned}$$

5.2 Experiments

For the final experiments, we use the full-length *Don Giovanni* performance conducted by Ádam Fischer to test the various combination strategies. In total, seven strategies are tested and evaluated on the 5,304 bar-level annotations. For each of these, we report the mean error and cumulative percentages for three error bounds.

Table 5 shows the results. Clearly, all combinations outperform the baseline, with the mean error being more than halved. The type of mean used to smooth the music/speech detector predictions does not have much influence. However, the combined music and speech-based fusion strategy outperforms all of the other methods. The best results are obtained with a constant average for the weights, with a mean error of 373 ms and an increase of the different percentages by more than 1.6% points compared to the baseline. We thus select this strategy in the following.

Table 5 Tracking errors on full-length *Don Giovanni* conducted by Á. Fischer, with different strategies to combine the two parallel trackers

Strategy	Mean type	Mean (ms)	≤1 s (%)	≤2 s (%)	≤5 s (%)
Baseline		811	91.8	95.0	97.3%
Late fusion		384	93.1%	96.6%	99.0 %
Speech fusion	Constant	392	92.9%	96.7	98.9
	Weighted	384	93.0	96.7	99.0
Music fusion	Constant	402	92.9	96.5	98.9
	Weighted	398	92.9	96.5	98.9
M and S fusion	Constant	373	**93.4**	**96.8**	**99.0**
	Weighted	374	93.3	**96.8**	99.0

Table 6 Error on full-length *Don Giovanni* conducted by Manacorda, using the combination of two parallel trackers

Strategy	Mean	≤1 s (%)	≤2 s (%)	≤5 s (%)
Baseline	561	90.1	94.5	97.9
M and S fusion	**547**	**90.3**	**94.7**	**98.0**

As a final test, we apply the combined tracker to our independent test piece–the Manacorda performance (using the Karajan as the score reference, as always), and hope to see similar improvement as in Table 5.

The result given in Table 6 is somewhat sobering: the improvement brought about by the music and speech strategy is minimal. As the alignment results on the recitative-only part did show a substantial improvement on the Manacorda (see Table 4), we conclude that it is not the recitative-specific features that prevent an overall improvement, but that it must be the combination strategy, or the audio classifiers (music/speech) controlling it. Recitatives are sonically extremely diverse in different recordings, especially regarding singing and speaking style. One immediate goal of future research will be to perform broader and more detailed investigations into classifiers that can reliably distinguish aria-style singing from recitative-style singing, which will, first of all, again require curating a broader set of annotated opera data.

6 Discussion and Conclusion

This paper has focused on the problems encountered by a state-of-the-art opera tracking algorithm when faced with *recitativo* sections. We isolated these sections in two live recordings of a full-length typical opera and tuned specific features in order to improve the accuracy of the alignment. Then, we proposed a new method for

opera tracking that involves two trackers working in parallel–one relying on speech-sensitive features, the other on music-sensitive features–and a way of combining these based on online music and speech classifiers. We validated the proposed method on a complete unseen opera, with (mostly) encouraging results. However, there remain a number of open problems.

The results reported in Tables 3 and 4 show that the optimized features help to increase by 8 and 10% the number of bars detected with an error less than one second in the recitative corpus, to a level of about 70%. This is still not at the same level of precision at which we can track non-recitative passages (cf. Table 2), and may not be sufficient for all kinds of desirable applications (such as automated subtitle display in the opera hall, or live video cutting for live streaming). In describing the numerous challenges arising in recitative tracking, we emphasized that the only reliable invariant among different performances is the linguistic content. Thus, audio-to-lyrics alignment would be a promising direction to pursue, but this implies new challenges. There is no freely available opera corpus for training appropriate phoneme recognition models; such training corpora would likely have to be singing- and language-specific. For instance, available Italian corpora, such as MLSPKA, CommonVoice, or Damp Dataset, involve only speech signals and karaoke songs. Thus, we will again have to curate our own training data; as a next step, we plan to annotate full-length operas at the sentence level in order to train such audio-to-lyrics models.

Another problem, which was hinted at in Sect. 2, concerns *structural mismatches* (e.g., parts omitted or variants inserted) between score (reference) and performance (target). As a future task, we plan to address this problem by adapting existing methods (e.g., *JumpDTW* [17]) to a real-time scenario.

The present study was conducted on only one opera,[3] *Don Giovanni*, in three different versions. Over time, more operas from other composers, periods, and genres will need to be investigated. Annotating even one full opera involves a lot of work–in our case involving more than 15,000 bar-level annotations over three performances–and it would have been nice to provide the resulting dataset to the scientific community. Unfortunately, the rights to our *Don Giovanni* recordings are with the Vienna State Opera. However, we have recently been involved in an amateur recording of Mozart's complete *Zauberflöte*, which we are in the process of annotating (including lyrics), and which will be made available to the community in the near future.

Acknowledgements This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowsa-Curie grant agreement No. 765068 (Project MIP-Frontiers). Special thanks to Christopher Widauer for providing the *Don Giovanni* recordings from the Vienna State Opera.

[3]We are aware of the fact that a single opera (in three different recordings) may seem like rather little data. However, it should be kept in mind that a full opera represents several hours of music and requires massive annotation work. For instance, the score of *Don Giovanni* is a book of 500 pages, containing 5304 bars of music.

References

1. Dannenberg RB (1984) An on-line algorithm for real-time accompaniment. In: Proceeding of the international computer music conference (ICMC). France, Paris, pp 193–198
2. Arzt A, Widmer G (2015) Real-time music tracking using multiple performances as a reference. In: Proceeding of the international society for music information retrieval conference (ISMIR). Málaga, Spain, pp 357–363
3. Arzt A, Widmer G, Dixon S (2008) Automatic page turning for musicians via real-time machine listening. In: Proceeding of the European conference on artificial intelligence (ECAI). Patras, Greece, pp 241–245
4. Lartillot O, Cancino-Chacón C, Brazier C (2020) Real-time visualisation of fugue played by a string quartet. Proceeding of the sound and music computing conference (SMC). Italy, Turin, pp 115–122
5. Arzt A, Frostel H, Gadermaier T, Gasser M, Grachten M, Widmer G (2015) Artificial intelligence in the concertgebouw. In: Proceeding of the international joint conference on artificial intelligence (IJCAI). Buenos Aires, Argentina, pp 2424–2430
6. Matthew P, David G, Alex H, Youngmoo KE (2013) Orchestral performance companion: using real-time audio to score alignment. IEEE MultiMedia 20(2):52–60
7. Brazier C, Widmer G (2020) Towards reliable real-time opera tracking: combining alignment with audio event detectors to increase robustness. In: Proceeding of the sound and music computing conference (SMC). Italy, Turin, pp 371–377
8. Dixon S (2005) An on-line time warping algorithm for tracking musical performances. In: Proceeding of international joint conference on artificial intelligence (IJCAI). Edinburgh, Scotland, UK, pp 1727–1728
9. Gadermaier T, Widmer G (219) A study of annotation and alignment Accuracy for performance comparison in complex orchestral music. In: Proceeding of the international society for music information retrieval conference (ISMIR), Delft, The Netherlands
10. Arshia C, Diemo S, Norbert S, Christopher R (2007) Evaluation of real-time audio-to-score alignment. In: International symposium on music information retrieval (ISMIR). Austria, Vienna, pp 315–316
11. Maëva G, Nathalie H, John S, Joe W (2010) Vocal tract adjustments in the high soprano range. J Acoust Soc Am 127(6):3771–3780
12. Villavicencio F, Bonada J, Yamagish J, Pucher M (2015) Efficient pitch estimation on natural opera-singing by a spectral correlation based strategy. In: Technical report, information processing society of Japan (IPSJ)
13. Johan S (1977) The acoustics of the singing voice. Sci Am 236(3):82–91
14. Gong R, Cuvillier P, Obin N, Cont A (2015) Real-time audio-to-score alignment of singing voice based on melody and lyric information. International speech communication association (INTERSPEECH). Dresden, Germany, pp 3312–3316
15. Vijayan K, Gao X, Li H (2018) Analysis of speech and singing signals for temporal alignment. Asia-Pacific signal and information processing association (APSIPA). Honolulu, HI, USA, pp 1893–1898
16. Majeed SA, Husain H, Samad SA, Idbeaa TF (2015) Mel frequency cepstral coefficients (MFCC) feature extraction enhancement in the application of speech recognition: a comparison study. J Theoret Appl Inf Technol JATIT) 79(1):38–56
17. Fremerey C, Müller M, Clausen M (2010) Handling repeats and jumps in score-performance synchronization. In: Proceeding of the international society for music information retrieval conference (ISMIR). The Netherlands, Utrecht, pp 243–248

Music Information Retrieval

Classification, Recommender System, Clustering, Pattern Analysis, Artificial Intelligence

IFSC: A Database for Indian Folk Songs Classification

Anshul Patel, Anuj Shah, Krutarth Gor, and Sapan H. Mankad

Abstract India is known for its culture of traditional folk music. Folk music represents the identity of different Indian regions and festival celebrations by people in those regions. Many popular Bollywood songs are based on folk music. There is a wide scope to apply artificial intelligence in this domain. However, to the best of our knowledge, there has been no availability of any audio dataset which represents Indian folk songs from various regions. In this work, we introduce a novel audio dataset of 307 folk songs from five major regions of India, namely Assamese, Marathi, Kashmiri, Kannada and Uttarakhandi. A pre-trained ResNet-based model has been used on mel-spectrogram-based audio representations for classifying a given folk song into one of these five regions. Results indicate that mel-spectrogram-based representation provides better performance in comparison with traditional spectrogram and short-term spectral feature-based representations.

Keywords Folk songs classification · Spectrograms · Music database

1 Introduction

Folk songs represent traditional culture of a geographic region. Different regions have their own regional identity through their folk songs for a diversified country like India. Use of instruments and rhythm in folk songs exhibits a specific pattern which helps in identifying the region from where a folk song originates. A machine learning algorithm may be trained to automatically detect the region of folk songs. There

A. Patel · A. Shah · K. Gor · S. H. Mankad (✉)
CSE Department, Institute of Technology, Nirma University, Ahmedabad, India
e-mail: sapanmankad@nirmauni.ac.in

A. Patel
e-mail: 16BIT016@nirmauni.ac.in

A. Shah
e-mail: 16BIT017@nirmauni.ac.in

K. Gor
e-mail: 14MCEN07@nirmauni.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_15

are a few publicly available databases such as Dunya under CompMusic project[1] and Saraga datasets [1] specifically for Indian music. However, there is a lack of standard databases in this area, especially in Indian sub-continent, and hence, it was felt that there is a need to manage a database of region-wise folk songs which may be helpful in several content-based music retrieval tasks such as region identification, instrument recognition and Raag identification.

There are 28 states and 9 union territories in India. They all have their own culture, music, life style and language. Some of the famous folk songs in India are as follows: Dandiya and Garba from Gujarat, Sambalpuri from Odisha, Lavni from Maharashtra, Bihugeet from Assam and so on. In Indian film and music industry, Bollywood has also played a key role in popularizing Indian folk songs. Due to diversity, each region has different pattern of lyrics and tune in its folk songs, thereby making it a difficult task for tourists and people of other regions to determine the native region of the song. This motivated us to prepare a database of folk songs and demonstrate our approach for folk songs classification using the database.

The remainder of this paper is as follows. In the next section, we discuss the related work carried out in this area. The folk songs dataset is introduced in Sect. 3. In Sect. 4, short-term spectral features extracted from audio files are discussed. Section 5 presents the experimental scenario and discusses the results at length. Finally, the last section concludes the paper with our observations.

2 Related Work

Most of the tasks for music information retrieval (MIR) are directed toward genre-based audio classification, and this work is mostly done for the western music. However, no significant work has been done specially in the area of folk song classification. In the literature, song classification has taken place mainly using two approaches: (i) acoustic features and (ii) symbolic features. Further, songs are classified using two types: (i) genre-based approach [2–4], and (ii) region-based approach [5]. There have been some attempts of pattern discovery from folk songs [6–8], music era classification [9], carnatic music and Raag identification [10–14]. They represented melody of an audio signal by the pitch of the predominant melodic source, which was estimated in carnatic music using the method proposed in [15]. This method performed favorably in music information retrieval evaluation exchange (MIREX 2011) on a variety of music genres. Majority of research has been carried out for Chinese, Greek and Austrian folk songs.

In [5], authors have carried out implementation on input audio files of WAV and MP3 formats with the dataset of 495 songs (328 for training and 167 for testing). They extracted 12 basic acoustic features and used support vector machine (SVM) classifier by combining it with four feature selection techniques. A symbolic musical feature density (MFDmap) feature was proposed in [16] to represent Chinese folk

[1] https://dunya.compmusic.upf.edu/.

Table 1 Summary of the work in folk songs classification

Paper	Region	Dataset size (# files)	Features	Classifier
[5]	Chinese	495	Acoustic	SVM
[17]	Chinese	312	Symbolic MFDMap	ELM
[18]	German, Austrian	210	Symbolic MFDMap	ELM
[19]	Greek	192	MFCC	Deep CCA
[23]	Irish, German, Austrian	491	Melody	Hidden Markov model
[20]	German, Chinese, Swedish	–	Musical motif	Attention model
[24]	Chinese	1329	Acoustic	SVM

songs. Symbolic notations in form of 16 bit monophonic audio wav file were used in [17] with a corpus of size 312 songs (281 for training, 31 for testing) using extreme learning machine (ELM)-based approach. MFDMap features were used to encode symbolic features of music. 106 German and 104 Austrian folk songs (168 for training, 42 for testing) were used in [18] with symbolic notations (16 bit monophonic audio wav file), again using MFDMap features. Further, 98 songs from Pontus and 94 from Asia Minor (in form of monophonic wav audio file) were used in [19] in which a 28-dimensional MFCC feature vector was used for each frame of the size 30 ms for feature extraction, and classification was carried out using deep canonical correlation analysis (CCA), and performance was measured with the auditory spectro-temporal modulation.

An attention-based architecture has been exploited in [20] in an unsupervised manner using musical motif embeddings to represent folk songs from Germany, China and Sweden. Conditional random fields (CRF)-based folk songs model has been proposed in [21, 22] for regional style classification of Chinese folk songs.

To the best of our knowledge, there is no publicly available dataset for folk songs in Indian languages. This motivated us to contribute a database using Indian folk songs. We restrict ourselves to only acoustic features in this work and present our analysis using short-term spectral and spectrogram-based features. We also implement a deep learning-based classifier for identifying region of a particular folk song. Table 1 summarizes the existing literature studied in this work.

3 The IFSC5 Corpus

In this section, we introduce a novel folk songs database of 307 songs, referred to as Indian Folk Songs Corpus (IFSC5). From different parts of the country, we selected five major regions, and songs are distributed into five categories describing their

Table 2 The IFSC5 corpus: the number of songs and 5 s clips within each class

Part	Region	# Songs	# Clips	Relative frequency (%)
Northeastern	Assamese	141	798	44.1
Northern	Uttarakhandi	29	174	9.6
Northern	Kashmiri	39	241	13.3
Southern	Kannada	63	384	21.2
Western	Marathi	35	210	11.6
Total		307	1807	

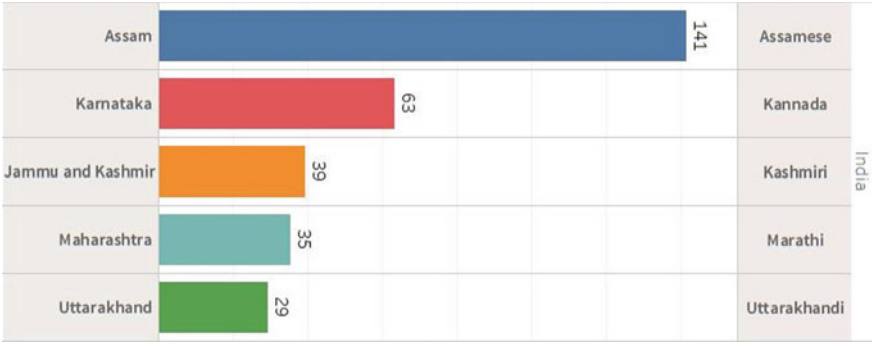


Fig. 1 Distribution of songs from the dataset

origin. This includes class-wise distribution of songs as depicted in Table 2, namely Assamese, Uttarakhandi (Pahadi), Marathi, Kashmiri and Kannada folk songs. Songs are collected from different publicly available websites, other online sources and commercially available recordings. The classification task is to predict the origin for a given song. All the songs were collected in wav format with the sampling rate of 44.1kHz. Since the songs were collected from different sources, they are differing in bitrate quality ranging from 352 kbps to 2822 kbps. We have published this dataset on GitHub repository [2] for academic and research purpose. Class-wise distribution of these songs is represented in Fig. 1. All the songs are of duration 4 to 5 min. We performed manual editing so as to extract clips of 5 s from each audio file. In contrast to [17] wherein authors used first 5 s chunks of audio, we used 5 s chunks from entire audio. Thus, multiple clips of 5 s duration were generated from every song. Zero score normalization was applied on every frame. Apart from this, no additional preprocessing such as vocal and instrumental separation or silence removal was carried out, and the features were extracted from audio comprising of both vocal and instrumental information. Mel-spectrogram features were extracted from this 5 s long clips, which are shown as an illustration for Assamese region in Fig. 6 along with the MFCC feature visual representation of the audio clip.

[2] https://github.com/anuj200199/A-dataset-of-Indian-Folk-Songs.

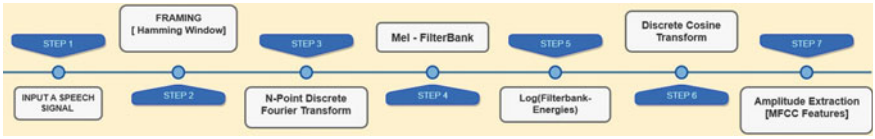


Fig. 2 MFCC feature extraction process

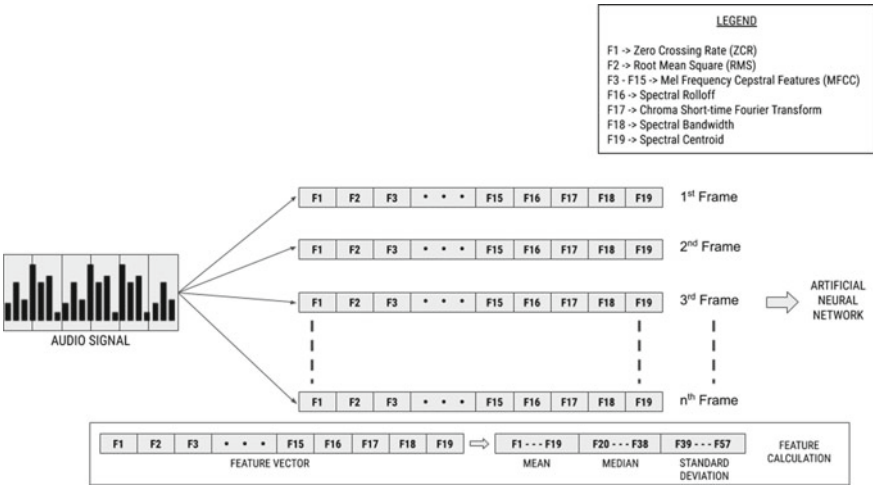


Fig. 3 Short-term spectral features extraction

4 Feature Representation

We used short-term spectral features for our experiments to represent the data. Mel-frequency cepstral coefficients (MFCC) is one of the most widely used feature for representing audio data [25]. The methodology for extracting MFCC features is specified in Fig. 2. Spectrograms were constructed from code given in this blog.[3] Librosa package of Python with default parameter values was used to extract features and Mel spectrogram from clips of audio.

Our objective in this work was to observe the impact of short-term spectral features over folk songs audio files. Features used in this work are given in Fig. 3. A 19-dimensional feature vector was used to represent each frame of the audio clip as depicted in Fig. 3.

[3] http://www.frank-zalkow.de/en/code-snippets/create-audio-spectrograms-with-python.html?i=1.

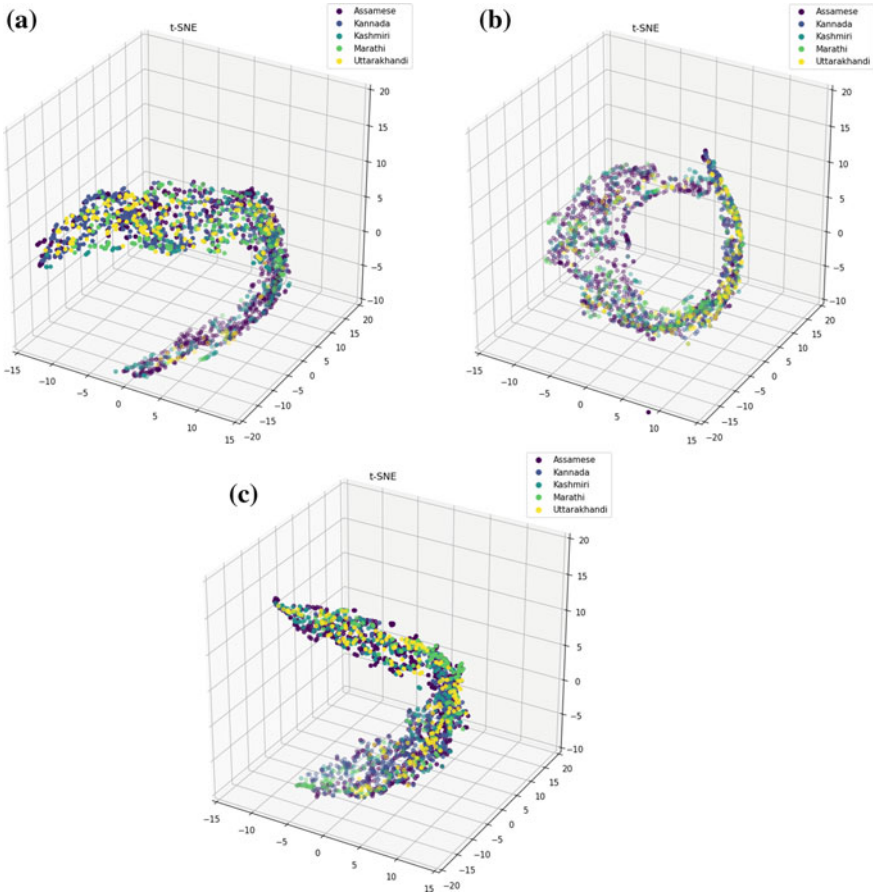


Fig. 4 t-SNE-based visualization of **a** mean **b** median **c** standard deviation features derived from original feature set

Measures of central tendency (mean, median and standard deviation) were computed from these feature vectors, thus forming another set of 19-dimensional feature vector for each case. A t-SNE-based 3D visualization of these features is depicted in Fig. 4 to show feature distribution among the classes.

5 Experiments

In this section, we describe our proposed methodology for classifying folk songs, the classification strategy, evaluation parameters and result analysis.

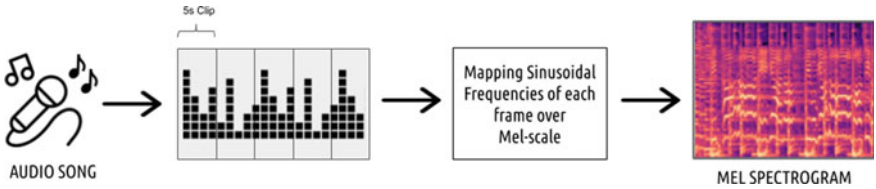


Fig. 5 Mel-spectrogram extraction process

5.1 Methodology

We extracted mel-spectrogram features from each edited audio file. The process of extracting mel-spectrograms is depicted in Fig. 5. Every audio file was clipped to form 5 s chunks and was treated independently. The mel-spectrogram image representation had a dimension 432 × 288 and was input to the classifier. Following systems are implemented in this work:

1. The 19-dimensional mean, median and std. dev. vectors were passed to artificial neural network. Early fusion of them was also carried out.
2. Audio to MFCC visualization as image and feed to pre-trained ResNet for classification.
3. 13-dimensional MFCC features were presented to ConvNet.
4. Each song clip was represented by spectrogram and fed to pre-trained ResNet for classification.
5. Audio to mel-spectrogram and fed to pre-trained ResNet for classification.

5.2 Classifier

Residual network (ResNet) is a deep architecture with 150+ layered neural network. The model became successful because of its capability of overcoming vanishing gradient problem, by introducing skip connection between layers. Different visual representations of the audio clip were used as input to a convolutional neural network built using architecture described in Fig. 7.

ResNet uses the fact that if normalizing the input helps to learn more features as well as reduce cost of computation, then a similar approach can be used after hidden layers. As a result, after some hidden layers, batch normalization is performed on the hidden layer output to improve training. We used the pre-trained ResNet50 model from Keras library which has pre-trained weights of ImageNet. We then fine-tuned the model for our set of training images. The model was trained for 10 epochs with batch size of 32, learning rate = 0.001 with decay = 0.005.

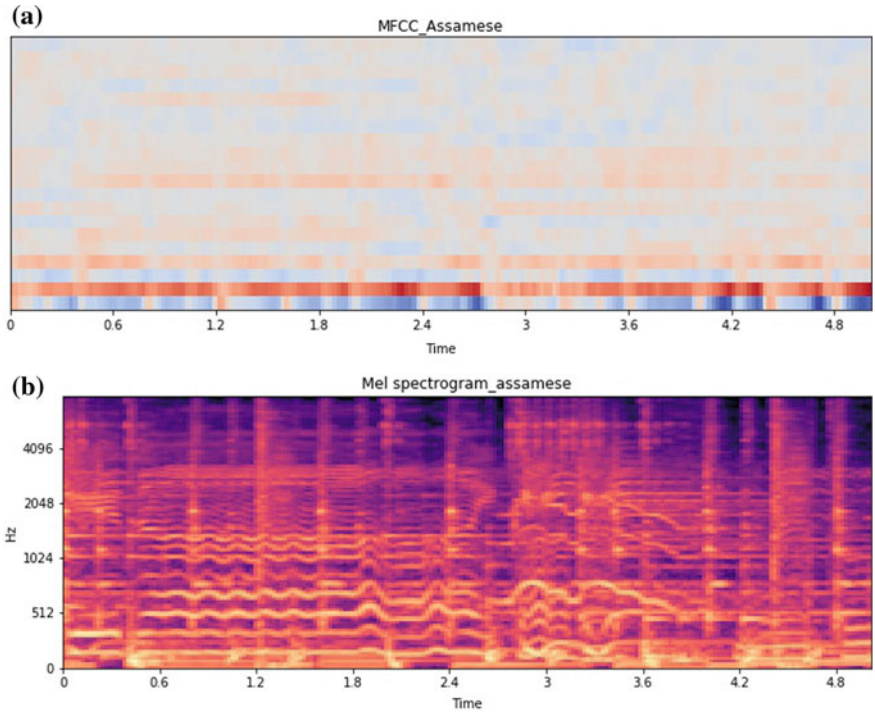


Fig. 6 **a** MFCC visual representation for assamese song **b** Mel-spectrogram visualization for assamese song

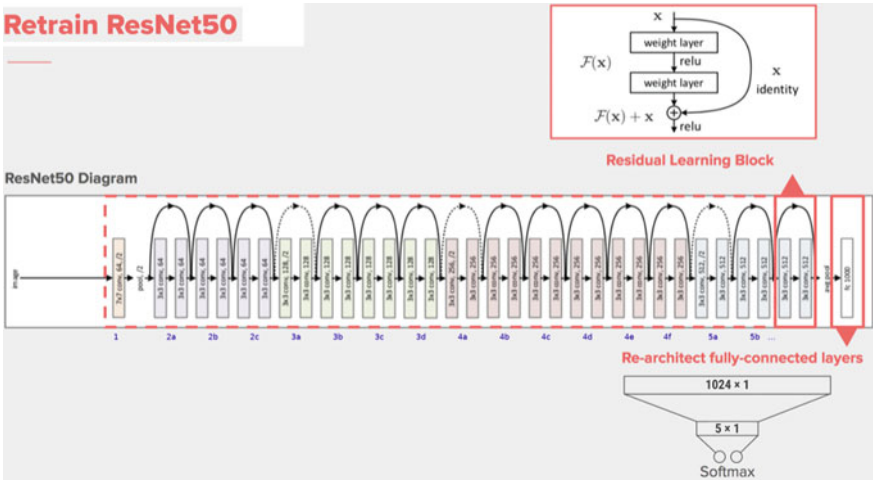


Fig. 7 Architecture of ResNet used in this work (adapted from: https://i.stack.imgur.com/gI4zT.png)

5.3 Performance Evaluation

Usually, in the literature for folk songs classification, accuracy-based comparison has been carried out. However, due to imbalance nature of this database, in order to verify the performance on more than one evaluation parameters, we also calculated following measures: (a) precision, (b) recall and (c) F-score. Table 3 shows the partition of dataset for evaluation purpose.

For a test sample of some particular class, following variables are defined:

- *True Positive (TP)*: Number of test samples correctly classified as positive with respect to a particular class
- *False Positive (FP)*: Number of test samples incorrectly classified as positive with respect to a particular class
- *True Negative (TN)*: Number of test samples correctly classified as negative with respect to a particular class
- *False Negative (FN)*: Number of test samples incorrectly classified as negative with respect to a particular class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Often, the harmonic mean of precision and recall value is also taken into account for evaluating the classifier performance.

$$\text{F score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The computation of macro-precision, recall and F-score is done on the basis of average values of their corresponding values from different trials of experiments. Usually, such measures are applicable when the underlying problem is a multiclass classification task with an assumption of uniform weights. The weighted average

Table 3 Partition of dataset in terms of number of clips

Region	Train set	Test set	Total
Assamese	599	199	798
Kannada	288	96	384
Kashmiri	181	60	241
Marathi	158	52	210
Uttarakhandi	131	43	174
Total	1357	450	1807

precision and recall values are calculated by using the number of samples of each class during calculation. Since the dataset is not purely balanced, we have reported the results using other parameters also in addition to classification accuracy. Multiple evaluation measures were calculated to compare different systems implemented in this work.

5.4 Results and Discussion

Experiments were done on spectrogram and mel-spectrogram representations apart from statistical feature vectors. Results are given in Tables 4 and 5. First, experiments results show that median features perform better than mean and std. dev. features in terms of all evaluation measures. We then tried with fusion of these features and found out that fusion does improve the performance of the system to some extent.

Spectrographic representation of audio does not yield significant performance (67.5%), and it performed worse than mean (73.1%) and median (74.6%) based statistical features. While representing the audio using mel-spectrogram, it boosted the accuracy to 89.7% thereby indicating that this audio representation is promising for such tasks. The mel-spectrogram-based system outperformed rest of the systems in terms of other parameters as well. This comparison is depicted in Figs. 8 and 9.

From the results in Tables 4 and 5, it can be seen that mel-spectrogram-based ResNet model outperforms other approaches. Looking at the graphs, it is evident that it is difficult to identify Uttarakhandi folk songs compared to others, whereas Assamese are easiest to identify. However, this may be the case due to imbalanced nature of the class. Assamese, being the class with highest number of samples (44%), and Uttarakhandi, being the lowest number of samples(10%), may mislead the classifier to produce biased results. In a nutshell, mel-spectrogram performs in reasonable way, and others are not able to exhibit good performance.

6 Conclusion and Future Directions

In this paper, we demonstrated a region-based folk songs classification on newly proposed dataset of 307 folk songs from five regions of Indian territory. We compared performance of traditional classifiers using short-term spectral features with a pre-trained ResNet-based deep classifier on the newly introduced IFSC5 corpus of Indian folk songs. We found that deep learning-based system with mel-spectrogram audio representations as input outperforms all other audio representations. This indicates that there is a lot of scope in deep architectures to improve the existing state of the art in music information retrieval based other tasks as well.

Table 4 Results on ANN classifier

	Accuracy		Macro-precision		Macro-recall		Macro-F-score		Weighted precision		Weighted recall		Weighted F-score	
	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set
Mean vector (19-d)	80.6	73.1	80.2	75.6	71.7	62.5	67.9	59	82.8	76.3	80.5	72.9	77.5	69.6
Median vector (19-d)	80.4	74.6	63.5	58.6	70.2	63.2	67	60.6	73.2	67.9	80.4	74.6	76.5	70.9
Std. dev. vector (19-d)	66.8	59.5	37.9	33.9	45.9	41.1	40.6	36.2	53.9	48.5	66.7	59.5	58.8	52.6
Mean + median vector (38-d)	90.3	74.2	88.4	63.7	82.6	63.3	82.3	62.3	90.6	70.6	90.2	74.2	88.7	71.6
Mean + std. dev. vector (38-d)	94.9	79.5	93.7	71	90.8	69.9	91.5	69.8	95.3	78.2	94.9	79.5	94.6	78.5
Median + std. dev. vector (38-d)	98.7	78.6	98.3	74	98.2	73.8	98.2	73.7	98.7	79.7	98.6	78.6	98.6	79
Mean + median + std. dev. vector (57-d)	96.1	77.1	94.7	70.4	93.0	70.1	93.3	70.1	96.5	77	96.1	77	97.4	76.8

Table 5 Results on pre-trained CNN classifier

	Accuracy		Macro-precision		Macro-recall		Macro-F-score		Weighted precision		Weighted recall		Weighted F-Score	
	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set	Train set	Test set
MFCC visualization to ResNet	76.5	65.5	81.3	51.4	67.7	54.9	67.1	53	78.6	59.3	77	65.4	74.1	62.1
MFCC coefficients to ConvNet	87.4	80.2	86	77.2	83.4	72.7	83.4	73	88.9	81.8	87.4	80.1	87.2	79.4
Audio to spectrogram and to ResNet	89.9	67.5	90.8	73.4	84.9	55.4	87.3	59.6	90.5	71.7	89.8	67.5	89.9	66.3
Audio to mel-spectrogram and to ResNet	98.4	89.7	97.8	89	98.2	87.9	97.9	88.2	98.4	89.5	98.3	89.7	98.3	89.6

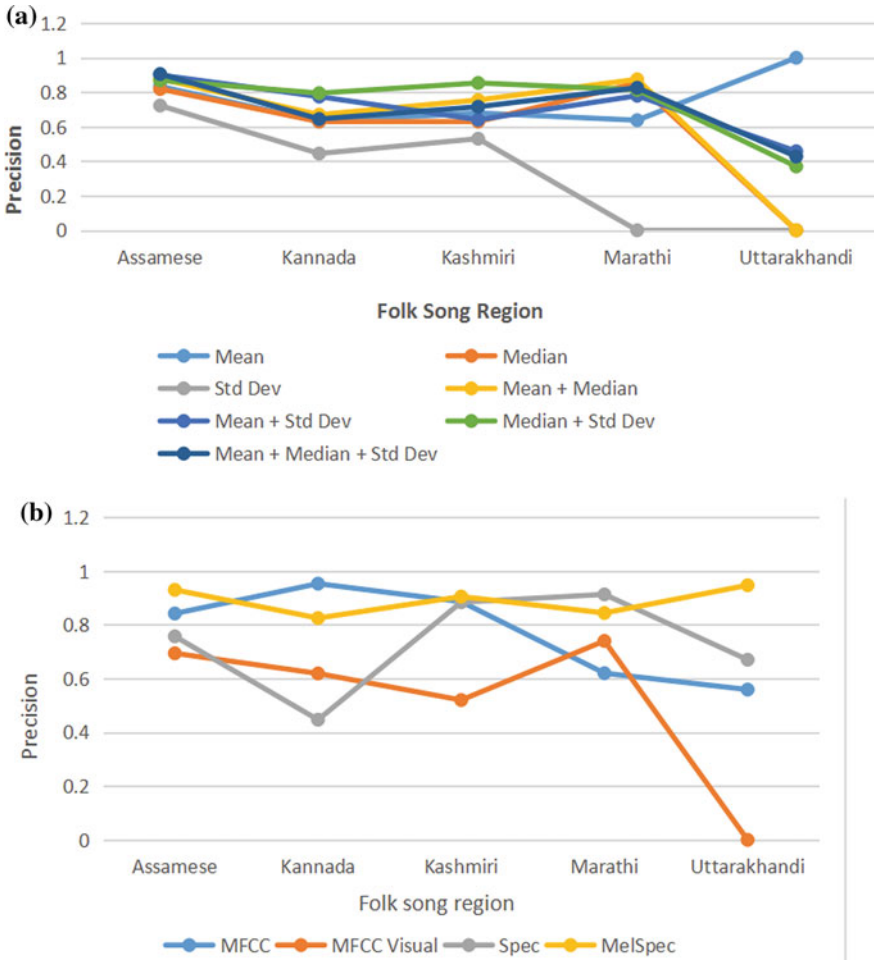


Fig. 8 Precision values for different classes using **a** statistical features on ANN architecture **b** visual and spectrographic representations on CNN architecture

In the future, improvement can be brought to existing performance by pre-processing the audio files further. This can be achieved by first separating the vocal and instrumental parts, and then, treating them differently may help in a better way. This dataset can be used in future for experimenting the impact of instruments in determining the folk song regions or even language identification. Further, methods of handling class imbalance can also be applied on this dataset.

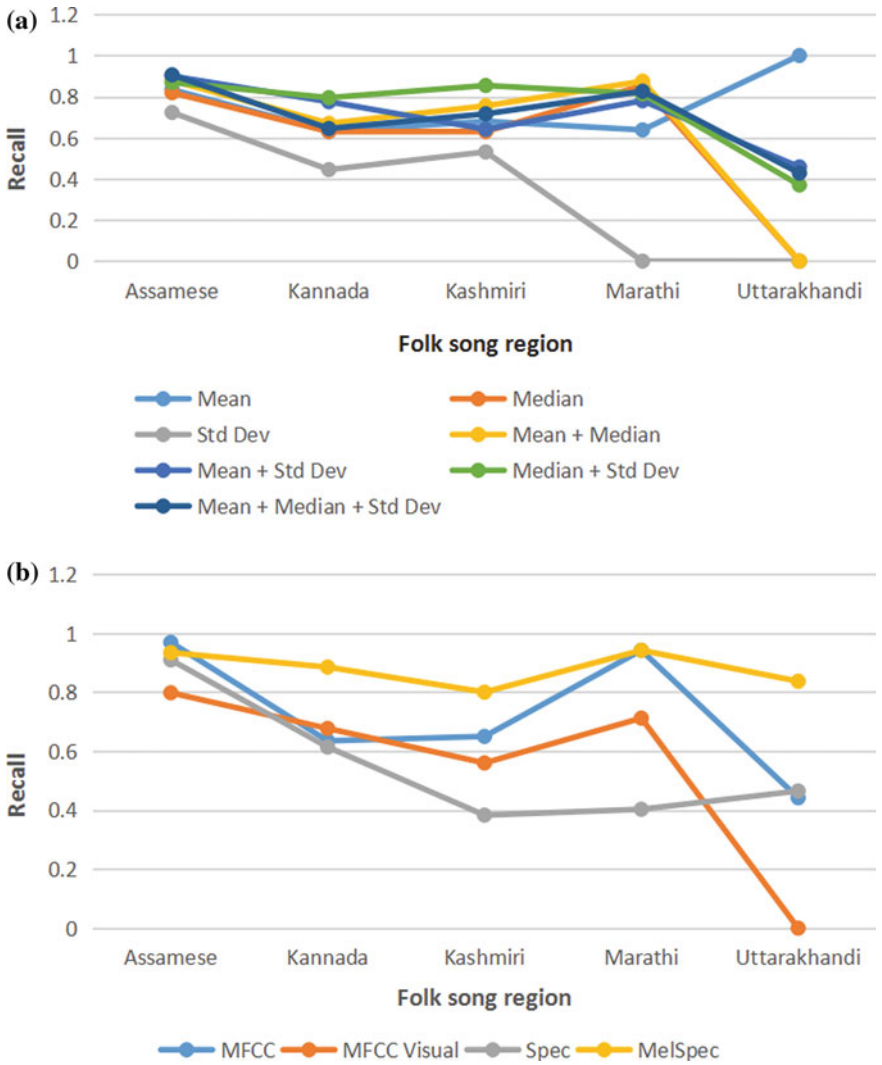


Fig. 9 Recall values for different classes using **a** statistical features on ANN architecture **b** visual and spectrographic representations on CNN architecture

Acknowledgements Authors acknowledge the insights provided by several people during discussions on this folk music research. We also thank to reviewers for their comments to improve the paper.

References

1. Arronte-Alvarez A, Gomez-Martin F (2019) An attentional neural network architecture for folk song classification. *arXiv e-prints*, page arXiv:1904.11074
2. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Trans Speech Audio Proc 10(5):293–302
3. Bassiou N et al (2015) Greek folk music classification into two genres using lyrics and audio via canonical correlation analysis. In: International symposium on image and signal processing and analysis, pp 238–243
4. Bergstra J, Casagrande N, Erhan D, Eck D, Kegl B (2006) Aggregate features and adaboost for music classification. Mach Learn pp 473–484
5. Boot P, Volk A, De Bas W, Haas (2016) Evaluating the role of repeated patterns in folk song classification and compression. J New Music Res 45(3):223–238
6. Bozkurt B, Srinivasamurthy A, Gulati S, Serra X (2018) Saraga: research datasets of Indian art music
7. Boot P, Volk A, da Haas WB (2016) Evaluating the role of repeated patterns in folk song classification and compression. J New Music Res 45(3):223–238
8. Chowdhuri S (2019) Phononet: multi-stage deep neural networks for raga identification in hindustani classical music. In: Proceedings of the 2019 on international conference on multimedia retrieval, pp 197–201
9. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoustics Speech and Sig Proc 28(4):357–366
10. Gulati S, Serrà Julià J, Ganguli KK, Sentürk S, Serra X (2016) Time-delayed melody surfaces for rāga recognition. In: Devaney J, Mandel MI, Turnbull D, Tzanetakis G (eds) ISMIR 2016. Proceedings of the 17th international society for music information retrieval conference; 2016 Aug 7-11, New York City (NY).[Canada]: ISMIR; 2016. pp 751–7. International Society for Music Information Retrieval (ISMIR), 2016
11. Heshi R, Suma SM, Koolagudi SG, Bhandari S, Rao KS (2016) Rhythm and timbre analysis for carnatic music processing. In: Nagar A, Mohapatra DP, Chaki N (eds) Proceedings of 3rd International conference on advanced computing, networking and informatics. Springer India, New Delhi, pp 603–609
12. Khoo SS (2013) The single hidden layer neural network based classifiers for han chinese folk songs. PhD thesis, Swinburne University of Technology
13. Khoo S, Man Z, Cao Z (2012) Automatic han chinese folk song classification using the musical feature density map. In: 6th international conference on IEEE, pp 1–9
14. Khoo S, Man Z, Cao Z (2013) German versus austrian folk song classification. In: 8th IEEE conference on IEEE industrial electronics and applications (ICIEA), pp 131–136
15. Salamon J, Gomez E (2012) Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans Audio Speech Language Proc 20(6):1759–1770
16. Juan L, Jing L, Jianhang D, Xi Z, Xinyu Y (2019) Regional classification of chinese folk songs based on crf model. Multimedia Tools Appl 78(9):11563–11584
17. Liu Y, Wei L, Wang P (2009) Regional style automatic identification for chinese folk songs. In: 2009 WRI world congress on computer science and information engineering, vol 7, pp 5–9
18. Liu Y, Xu JP, Wei L, Tian Y (2007) The study of the classification of Chinese folk songs by regional style. In: International conference on semantic computing (ICSC 2007). IEEE, pp 657 – 662
19. Neubarth K, Conklin D (2016) Contrast pattern mining in folk music analysis. Comput Music Anal 393–424
20. Pandey G, Mishra C, Ipe P (2003) Tansen: a system for automatic raga identification. In: IICAI, pp 1350–1363
21. Peter Van Kranenburg AV, de Bruin M (2017) Documenting a song culture: the dutch song database as a resource for musicological research. Int J Digit Libr

22. Juan L, Jing L, Jianhang D, Xi Z, Xinyu Y (2019) Regional classification of chinese folk songs based on crf model. Multimedia Tools Appl 78(9):11563–11584
23. Salamon J, Gomez E (2012) Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans Audio Speech Language Proc 20(6):1759–1770
24. Sridhar R, Geetha TV (2008) Music information retrieval of carnatic songs based on carnatic music singer identification. In: International conference on computer and electrical engineering ICCEE 2008. IEEE, pp 407–411
25. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoustics Speech and Sig Proc 28(4):357–366

North Indian Classical Music Tabla Tala (Rhythm) Prediction System Using Machine Learning

Shambhavi Shete and Saurabh Deshmukh

Abstract Music information retrieval (MIR) and its applications have gained popularity after the advancements in machine learning techniques. In various applications such as genre detection, song recognition, automatic musical score generation, and music transcription, the identification of rhythm, tempo, stroke type, etc., is necessary. For Western Music, various applications have been developed to serve the purpose. However, very few researches have been done in North Indian classical Music (NICM). The rhythm instrument used in NICM called Tabla, and its rhythm patterns (Talas) are considered in this research. The system proposed here predicts the Tala (Rhythm). 'Timber' is a non-tangible audio attribute (descriptor) of sound. The descriptors that more clearly define a timber such as zero cross rate, roll-off, roughness, brightness, and irregularity) are combined with mel-frequency cepstral coefficient to extract the audio features. Overall, 366 audio excerpts containing sound samples of 9 basic Tabla strokes are used to train the system. Machine learning algorithms, multi-support vector machine, and feed-forward backpropagation neural network are used to identify the Tabla strokes, and frequency of occurrence of each basic stroke is calculated. The novel method proposed in this paper predicts the Tala with the help of identified strokes, the total number of strokes, and the frequency of occurrence of each stroke. The result shows that the Tala could successfully be predicted even if not all strokes from the Tala are correctly identified. Tala prediction accuracy of 93.33% and the Tabla stroke identification accuracy of 96.85% are obtained. The correlation factor of +0.77 between Tabla stroke identification and Tala prediction calculated proves that, even if not all Tabla strokes are identified correctly, the Tala could still be predicted using the basic structure of NICM's Tabla Tala.

Keywords Machine learning · North Indian classical music · Tabla · Tala (rhythm) · Timber

S. Shete
Career Point University, Kota, Rajasthan 324005, India

S. Deshmukh (✉)
Maharashtra Institute of Technology, Aurangabad, Maharashtra 431005, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_16

1 Introduction

Music information retrieval (MIR) is a rapidly growing branch of sound engineering that has various applications in the music industry, film, and television industry and in the fields of music recognition and synthesis. With the advancements in machine learning, music information retrieval has gained substantial popularity and interest across the world.

North Indian Classical Music (NICM) is one of the ancient classical music traditions in the world. Along with vocal recitation, in the NICM, different accompanying musical instruments such as Tanpura, Harmonium, Sarangi, Tabla, and Pakhawaj are used. The requirements and the theoretical structures of the use of accompanying musical instruments are explained in 'Natyashastra' [1]. This etiology explains the structure of the instrument, the material to be used, the tuning frequency of the instrument, and methods to play these accompanying musical instruments.

Tabla is one of such basic accompanying rhythm instrument used popularly by the vocalists during their vocal performances in NICM. Tabla is made up of two drums of different sizes and shapes. The left drum is metallic, hollow, dome-shaped, and tightly wrapped with animal skin. The right drum is made up of teak wood and smaller than the left drum. It is laced with hoops and wrapped with animal skin. Both the drums have dowels and hoops to tighten the drums. A paste mixture of rice flour and iron filings is applied on both the drum surfaces. This dried paste gives a defined pitch to the drums. This paste is applied to the skin in several layers [2]. The right drum is usually tuned with the musical note of the fundamental frequency of the singer's scale (C_4# for male singer and F_4# for female).

Tabla is also played in a solo form. Tabla player plays different forms and combinations of basic Tabla strokes. These combinations of basic Tabla strokes are performed by keeping the type of Tala (Rhythm) and tempo intact. The Tala is a rhythmic cycle of the beat. From the North Indian Classical Music Tala structure, it is observed that each basic Tala contains a fixed number of Tabla strokes that are dissimilar for different types of Tala. The basic Tabla Talas and their rhythm patterns with the frequency of occurrences of Tabla strokes are shown in Table 1. There are several methods to identify the stroke of a rhythm instrument. The percussion instrument used in music differs for the type of music. African music makes use of percussion instruments that are made up of animal skins. Western music and popular music make use of drums. Each of these instruments has different names of the rhythms and the basic strokes to perform the rhythm. The drum stroke identification [3] from polyphonic music requires separation of sound originating from different frequency bands dedicated to the part of the drum set being played. Apart from this, classification of the percussive sound [4], Automatic Transcription of Drum strokes [5, 6], Tonic Independent Mrudangam stroke Transcription [7], Akshara Transcription of Mrudangam [8], Tabla stroke Recognition and Resynthesize [9], and Polyphonic Musical Instrument Recognition [10] are some of the major researches carried out in music information retrieval domain, where the basic strokes of the instrument and rhythm patterns are identified. The typical process of identification of strokes from

Table 1 Basic Tabla Talas and rhythm pattern

Tala	Beats	Rhythm pattern	Tabla stroke frequency
Teental	16	Dha Dhin Dhin Dha | Dha Dhin Dhin Dha | Dha Tin Tin Ta | Ta Dhin Dhin Dha ||	Dha-6, Dhin-6, Tin-2, Ta-2
Dadra	6	Dha Dhin Na | Dha Tin Na ||	Dha-2, Dhin-1, Na-2, Tin-1
Rupak	7	Tin Tin Na | Dhin Na | Dhin Na ||	Tin-2, Na-3, Dhin-2
Jhaptal	10	Dhin Na | Dhin Dhin Na | Tin Na | Dhin Dhin Na ||	Dhin-5, Na-4, Tin-1
Kerva	8	Dha Ge Na Tin | Na Ka Dhin Na ||	Dha-1, Ge-1, Na-3, Tin-1, Ka-1, Dhin-1

rhythm instrument requires onset detection, audio segmentation, feature extraction, and classification. After detecting the basic strokes from the rhythm instrument, it is easy to recognize the rhythm pattern.

Similar to other world music, in NICM, the rhythm of the Tabla instrument is periodic. A Tala is a combination of a fixed number of total strokes, played in cycles. In this research, five basic Talas, namely Teental (16 strokes), Jhaptal (10 strokes), Rupak (7 strokes), Dadra (6 strokes), and Kerva (8 strokes) are considered. The selection of these Talas is based on their variety in the total number of beats as well as a variety of Tabla strokes being used to play the Talas. These Talas are selected that have a different number of total strokes and are some of the basic Tabla Talas used in NICM. The strokes of Tabla also have a pattern of repetition for some strokes, while some strokes are used only once in a cycle of the rhythm. The frequency of occurrence of Tabla strokes in the Talas could be counted by identifying the basic Tabla strokes. For example, Teentala is composed of 16 basic strokes, out of which 'Dha' and 'Dhin' strokes are repeated 6 times, and 'Tin' and 'Ta' strokes are repeated 2 times throughout the entire one cycle of the Tabla recording.

Prediction of Tala from the identified strokes is possible with the help of basic information provided to the system regarding the number of strokes in each Tala, types of strokes, and frequency of occurrence of these basic strokes in that Tala. For the learners of Tabla instruments, it is necessary to know the Tabla strokes that are being played by the Tabla performer. For the systems of automatic music recognition and transcription, it is important to identify the Tala. For the learned listeners/practitioners of NICM, it is a bit difficult to recognize the basic strokes and their repetition of the pattern to predict the Tala accurately. For common listeners, it is very challenging to recognize the Tala based on the basic Tabla strokes being played and the total number of repetitions of the strokes. Machine learning becomes a useful technology to train a system to automatically identify Tabla strokes, tempo, and predict the Tala.

In this research, we propose a North Indian Classical Music Tala (Rhythm) prediction system based on the type and number of Tabla strokes being identified using machine learning. The system uses a machine learning algorithm to identify the Tala

based on the detection of the Tabla stroke, frequency of occurrence of Tabla strokes, and a theory of the basic structure of a given Tala that formulates the combination of the stroke into the Tala.

2 State of the Art

Music information retrieval is a branch of sound information retrieval that deals with extracting meaningful musical information from a given piece of music. This musical information contains a low-level audio signal description. It is useful to identify a Tabla stroke. The descriptors that describe this low-level information of a sound are called low-level descriptors (LLD) [11]. There exist many digital signal processing (DSP) methods to extract musical information and represent it in various ways.

The features extracted from an audio play a prominent role in the identification of Tabla strokes. The audio features, also called audio descriptors, are categorized using different perspectives. The taxonomy of the audio descriptors differs with respect to their applications. There are global audio descriptors that are computed over the entire audio signal and instantaneous descriptors that are computed for each time frame. Temporal audio descriptors are computed to generate signal energy envelop, and spectral audio descriptors are computed by computing different transforms [11].

Sound is a physical phenomenon that is expressed in the form of pitch, amplitude, duration, and timber. Pitch is expressed in term of Hertz (Hz), amplitude is expressed in terms of decibels (dB), and duration is expressed in terms of seconds (s). However, other than these three measurable dimensions of sound, the timber is a multi-dimensional non-tangible entity. Timber cannot be expressed in terms of any unit. There are many ways with which researchers have explained the timber [12]. Timber is expressed using attack time, attack slope, zero crossing rate, roll-off, brightness, roughness, and irregularity, along with MFCC [13]. The zero cross rate is used to detect the number of times the signal crosses the x-axis. It is a measure of noise. Roll-off is used to represent the shape of a high-frequency energy signal. It is a frequency below which 85% energy signal is present. Brightness represents the spectral energy above the threshold frequency of 1500 Hz. Roughness is a sensory dissonance between all peaks of the spectrum of the signal. Irregularity is the degree of variation of the successive peaks of the spectrum of a signal. Mel-frequency cepstral coefficient (MFCC) is a set of the power spectrum which describes the shape of the sound [14].

It is very difficult to select appropriate audio descriptors that describe the timber of a sound accurately. There exist hundreds of audio descriptors and their variants. Hybrid selection method of wrapper approach promises to develop a subset of audio descriptors useful for given music information retrieval applications [15]. Research has proved that these audio descriptors are useful for singer identification [16, 17], gender recognition of a singer [18], Tabla stroke identification [19], and musical instrument identification [20].

There are many challenges to identify a Tabla stroke. In general, it is difficult to differentiate between the strokes 'Te' and 'Ti' without making use of a language model for each of the strokes. This is because the timbral descriptors of 'Te' and 'Ti' are much similar to each other, and a Tabla player plays these short duration strokes interchangeably. Also, a fall of accuracy for the identification of Tabla stroke is reported due to the mixture of the previous stroke's ringing sound with the current stroke. For example, Tabla stroke 'Na' could be misclassified due to the ringing sound of the previous stroke 'Dha', which is generated by hitting both the drums simultaneously. The 'Dha' is produced using 'Ga' (left drum) + 'Na' (right drum). Thus, it becomes difficult to separate this 'Na' with the next stroke 'Na' [21]. Tabla stroke identification is based on the overlapping of the Tabla strokes. Spectral centroid and ZCR features are useful to identify the power spectrum and overlapping between the Tabla strokes. The Tabla strokes 'Dha', 'Dhin', 'Din', and 'Tin' produce more overlapping due to the sustained sound.

3 The Method

An automatic Tabla stroke identification and Tala prediction system are shown in Fig. 1. The audio database is recorded through a professional Tabla player in a studio environment. Different Talas and individual basic Tabla strokes are recorded with multiple instances of each Tabla stroke. The Tabla instrument used is tuned to C_4# note. The database consists of the audio excerpts of 9 basic Tabla strokes with 25 samples of each stroke. The Tabla strokes considered here are 'Na', 'Ta', 'Te', 'Tun', 'Ga', 'Ka', 'Dha', 'Dhin, and 'Tin'. The Tabla strokes are digitized using a sampling frequency of 44.1 kHz with 16-bit PCM.wav file format.

The Tabla stroke location in an audio file is determined by identifying the onset locations. Onsets are the locations where the signal energy is high [22, 23]. Each audio excerpt is processed to identify the onsets. After identifying the onsets, the audio file is segmented. Each segment contains one Tabla stroke. The accuracy of the identification of the onsets and subsequent segmentation plays an important role in the Tabla stroke identification system. After the segmentation of all the audio excerpts, a total of 366 Tabla strokes is separated. Out of the total Tabla strokes segmented, 70% is used for training and 30% for testing.

Timbral audio descriptors, namely zero crossing rate (ZCR), roll-off, brightness, roughness, irregularity, and mel-frequency cepstral coefficients (MFCC), are extracted from the Tabla stroke samples [13].

We have considered two supervised learning algorithms for identification of the Tabla strokes, namely multi-support vector machine (multi-SVM) and feed-forward backpropagation neural network (FFBPNN). In general, the SVM is useful for binary classification; however, it could also be extended to be used for multi-class classification. In one versus one configuration, the output is binarized such that the target class represents 1 while other classes as 0. This configuration requires voting to identify

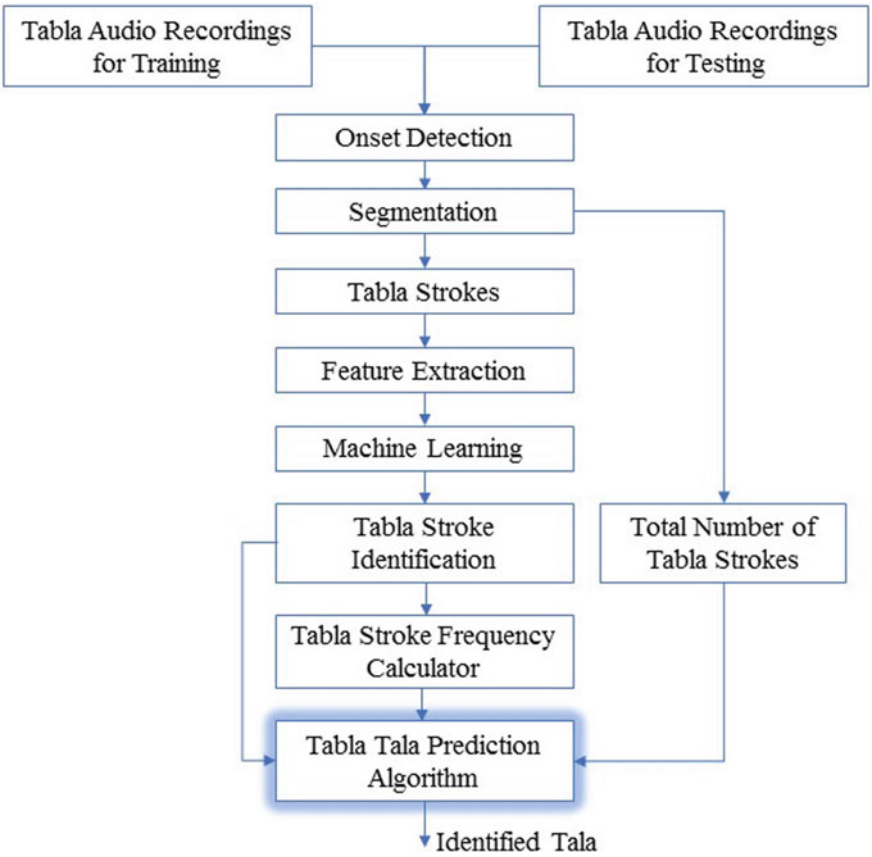


Fig. 1 Tabla stroke identification and Tala prediction system

a stroke that is a class [7]. Both machine learning algorithms are supervised and are useful in many music information retrieval applications.

Feed-forward backpropagation neural network (FFBPNN) has two passes, forward pass and backward pass [24]. The output of each neuron is calculated by the sum of the product of each input and its associated weights, followed by a squashing function. At the output, layer error is calculated by comparing the output and the target. In the backward pass, this error is then backpropagated from the output layer to the input layer in the form of a change of weights. Feed-forward backpropagation neural network (FFBPNN) is trained by 25 audio features extracted from audio excerpts.

Based on the types of Tabla strokes, the total number of Tabla strokes, and their frequency of occurrences, the Tala is predicted. The performance of the system is evaluated based on the accuracy of the identification of the Tabla strokes and Tala.

4 Experiments and Results

The experiments are carried for 9 basic Tabla strokes. The test samples are selected from 5 basic Tala. The audio features are categorized into two sets consisting of only MFCC and timber audio descriptors with MFCC. The feed-forward backpropagation neural network (FFBPNN) and multi-SVM are trained and tested to identify the Tabla strokes. When presented 15 test samples, 3 from each Tala, it is observed that FFBPNN supersedes in performance over multi-SVM for the identification of Tabla strokes. The automatic Tabla stroke identification system using FFBPNN gives a stable system than using multi-SVM. The results show that the generalization capability of a trained neural network accurately identifies Tabla strokes from unknown test samples whereas, being binary (one vs. one) classifier, multi-SVM does not guarantee a stable Tabla stroke identification system. Also, automatic Tabla stroke identification accuracy is highest for timber audio descriptors than MFCC when FFBPNN is used.

It is observed that the sound produced from 'Ga' is closed, and its sustain is small. Thus, the accuracy of identification of Tabla stroke 'Ga' is 100%. Timbral attributes of the sound produced by stroke 'Ga' are unique and have no match with any other Tabla stroke present in the database. On the other hand, for the Tabla stroke 'Tin', that generates overtones, merged with its fundamental frequency. From this mixed harmonic structure, it becomes difficult to correctly identify the Tabla stroke 'Tin'. The Tabla stroke 'Tin' is produced by hitting an open stroke on the right-hand drum. In contrast to this, when the Tabla stroke 'Na' is produced, the Tabla player damps the overtones produced after it. Thus, the 'Na' stroke has high onset energy while sharp sustain. We have compared the Tabla stroke originated from left, right, and both the drums with respect to their identification accuracy. It is observed that the right drums Tabla stroke identification accuracy is 96%, the left drum 84%, and for the strokes originated from both the drums simultaneously, it is 88%. The reason behind these differences could be explained with respect to the A (attack), D (decay), S (sustain), and R (release) of each stroke.

The method proposed here ensures that, even if the system does not detect all the Tabla strokes accurately, it is possible to predict the Tala to some extent. This is because few misclassified Tabla strokes do not affect Tala identification. In NICM, there exist Talas that contain the same number of total strokes, but the number of occurrences of the strokes and the locations of the strokes in the entire cycle of the Tala is different for different Talas. Thus, it is possible to predict the Tala even if one or more Tabla strokes are misclassified.

Figure 2 shows the results obtained for Tabla stroke identification and corresponding Tala Prediction. As shown in the table, the Tabla stroke identification accuracy for FFBPNN using timbral audio descriptors is 100%, and the corresponding Tala is also predicted correctly. However, for a few cases of Teental, even if not all Tabla strokes are correctly identified, the system predicts the Tala accurately. In the case of Rupak Tala, though, the Tabla stroke identification accuracy is 71.43% the Tala could not be predicted correctly. This is because the total number of basic Tabla

Tala	Test Audio	Accuracy (%)	Tala Predicted (Y/N)	Accuracy (%)	Tala Predicted (Y/N)	Accuracy (%)	Tala Predicted (Y/N)	Accuracy (%)	Tala Predicted (Y/N)
		FFBPNN				Multi-SVM			
		Timber		MFCC		Timber		MFCC	
TeenTal	T_1	87.5	Y	75	Y	81.25	Y	68.75	Y
	T_2	100	Y	100	Y	10	Y	93.75	Y
	T_3	93.75	Y	81.25	Y	81.25	N	81.25	N
Zaptal	Z_1	100	Y	100	Y	100	Y	90	N
	Z_2	100	Y	100	Y	100	Y	100	Y
	Z_3	100	Y	100	Y	100	Y	100	Y
Rupak	R_1	100	Y	85.71	Y	85.71	N	85.71	N
	R_2	100	Y	85.71	Y	71.43	N	71.43	N
	R_3	71.43	N	85.71	N	85.71	N	100	Y
Dadra	D_1	100	Y	50	N	100	Y	50	N
	D_2	100	Y	100	Y	100	Y	100	Y
	D_3	100	Y	100	Y	83.33	Y	100	Y
Kerava	K_1	100	Y	100	Y	75	Y	100	Y
	K_2	100	Y	85.71	Y	100	Y	100	Y
	K_3	100	Y	100	Y	87.5	Y	87.5	Y

Fig. 2 Tabla stroke identification and corresponding Tala prediction results

strokes identified is more than the basic strokes of Rupak Tala. In such a situation, the system could not predict Tala accurately.

When MFCC alone was used as an audio descriptor, the Tabla stroke identification accuracy has dropped to some extent. But still, the Tala could be predicted accurately. In this case also, for Rupak Tala, which has only three basic Tabla strokes, it is observed that, if the identified strokes belong to the basic Tabla strokes, the Tala could be predicted. However, if the identified stroke is not from the category of basic Tabla stroke of the given Tala, then the probability that the system may predict the Tala accurately would decrease. In Fig. 2, for test audio samples R_2 and R_3, the Tala could be predicted for the former while it could not be predicted for the latter, even though the Tabla stroke identification accuracy for both is 85.71%. When multi-SVM is used with MFCC and timber, the results were not promising, even though the Tabla stroke identification accuracy is high. This is due to the misclassification of the Tabla strokes into some other category.

Overall, the system performance is calculated with respect to the number of times the system could predict the Tala. The Tala prediction accuracy for FFBPNN using the timber audio descriptor is 93.33% and using multi-SVM 73.33%. Figure 2 shows the comparison of the performance of stroke identification and Tala prediction. It is observed that when the Tabla stroke identification accuracy is higher for FFBPNN, the Tala is predicted accurately. These results are similar for both timbral and MFCC audio descriptors. On the other hand, for multi-SVM, the stroke identification accuracy degrades the performance of Tala prediction. The correlation factor between

stroke identification and corresponding Tala identification is found to be +0.77, which indicates a strong uphill linear relationship between stroke identification and Tala predication. It is obvious that if the stroke identification accuracy is 100%, then the Tala prediction would also be accurate.

The Tala prediction algorithm proposed here works differently for different types of Talas. We have considered 5 basic Talas, and the logic could be extended to any number of Talas from NICM. The proposed algorithm considers three factors before predicting a Tala, viz. the total number of Tabla strokes, the number of basic Tabla strokes considered, and the frequency of each of the identified Tabla strokes. The proposed algorithm considers different logic for different types of Talas. As explained earlier in Table 1, for example, Teental consists of 4 basic strokes and 16 total numbers of strokes per cycle. The fixed frequency of occurrences of the strokes 'Dha' & 'Dhin' (6 times) and 'Ta' & 'Tin' (2 times) each ensures the presence of Teental. The algorithm predicts the Tala based on the fulfillment of the required frequency of strokes 'Dha' or 'Dhin'. In such cases, it is not necessary to identify all 16 strokes accurately. By checking the frequencies for these two strokes, the Tala could be predicted accurately. Similar logic applies to each Tala, and the logic differs for different Tala. The same algorithm can be further extended for any number of Tala present in North Indian Classical Music Tabla Talas (Fig. 3).

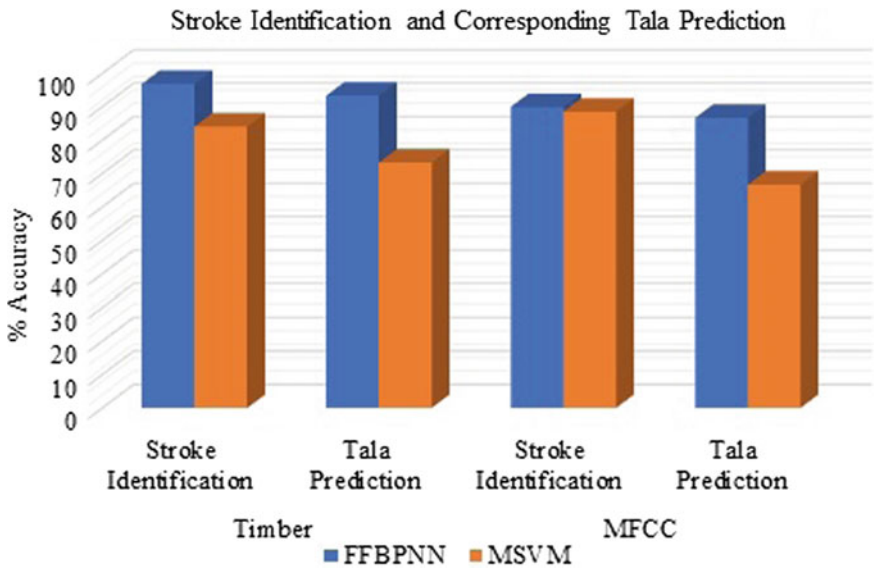


Fig. 3 Comparison of Tabla stroke identification and corresponding Tala prediction

5 Conclusion and Future Scope

Music information retrieval has gained substantial popularity in the music research domain. With the advancements of signal processing techniques and computational capabilities of neural networks, a lot of applications in the field of music have been invented. In this research, we have proposed a method for Tala prediction, based on the number and type of basic Tabla strokes identified. Total 9 Tabla strokes and 5 types of basic Tala are considered to train the system using FFBPNN and validated using multi-SVM. Traditional MFCC and non-tangible timbral attributes of sound are considered for feature extraction. For FFBPNN and multi-SVM, Tala could be predicted accurately when Tabla stroke identification accuracy was 100%. However, the proposed system in this research could also successfully identify Tala even though the Tabla stroke identification accuracy was lower.

The system proposes a method that uses North Indian Classical Music Tala structure in such a way that, for each Tala, if some basic strokes are correctly identified, then the Tala could be predicted accurately by using count and type of the basic strokes identified for that Tala. The Tabla stroke identification accuracy of the system is found to be 96.85% using FFBPNN, and the Tala prediction accuracy of 93.33%. The Tala prediction accuracy increases with the increase in Tabla stroke identification accuracy. The correlation factor of +0.77 between Tabla stroke identification and Tala prediction proves that accurate is the Tala prediction when Tabla stroke identification is accurate. The results show that Tala identification could be successfully achieved even for low Tabla stroke identification accuracy as 68.75%. In the future, the system and the proposed Tabla Tala prediction algorithm could be extended to any number of Tabla Tala in North Indian Classical Music. Also, the acceptable lower threshold for misclassified Tabla stokes that is capable of predicting Tala could be determined.

Bibliography

1. Beck G (2012) Sonic liturgy: ritual and music in Hindu tradition. University of South Carolina Press, South Carolina, 12 Dec 2012
2. Klippstein, Construction & Playing: Tabla, https://www.india-instruments.com/, 2012. (Online). Available: https://www.india-instruments.com/construction.html. Accessed 23 Mar 2020
3. Anssi K, Paulus J (2009) Drum sound detection in polyphonic music with hidden Markov models. EURASIP J Audio Speech Music Process 2009:9
4. Steelant DV, Tanghe K, Degroeve S, Baets B, Leman M, Martens J-P (2004) Classification of percussive sounds using support vector machines. In: Annual Machine Learning Conference of Belgium and The Netherlands (BENELEARN), Belgium and The Netherlands
5. Jacques C, Roebel A (2018) Automatic drum transcription with convolutional neural networks. In: 21st international conference on digital audio effects, Sept 2018, Aveiro, Portugal, September 2018
6. Vogl R (2018) Deep learning methods for drum transcription and drum pattern generation. Johannes Kepler University Linz, Linz

7. Anantapadmanabhan A, Bello J, Krishnan R, Murthy H (2014) Tonic independent stroke transcription of the Mrudangam. In: AES 53rd international conference, London, UK, 27–29 Jan 2014
8. Kuriakose J, Chaitanya K, Padi S, Murthy H, Umayalpuram S (2015) Akshara transcription of Mrudangam strokes in Carnatic Music. In: 2015 twenty-first national conference on communications (NCC), Mumbai, India
9. Chordia P, Ray A (2010) Tabla Gyan: an artificial tabla improviser, Jan 2010
10. Burred JJ, Robel A, Sikora T (2009) Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope. In: IEEE international conference on acoustics, speech and signal processing, Taipei
11. Geoffroy P (2004) A large set of audio features for sound description (similarity and classification) in the CUIDADO project, 23 Apr 2004
12. Park TH (2004) Towards automatic musical instrument timbre recognition, Candidacy
13. Lartillot O, MIR ToolBox Manual, Finnish Centre of Excellence in Interdisciplinary Music Research
14. Deolekar S, Siby A (2018) Tree-based classification of tabla strokes. Curr Sci 115:1724–1731
15. Deshmukh S (2012) A hybrid selection method of audio descriptors for singer identification in North Indian Classical Music. In: IEEE Explorer, Himeji, Japan, 2012
16. Deshmukh S, Bhirud S (2014) North Indian classical music's singer identification by timbre recognition using MIR toolbox. Int J Comp Appl 91(4):1–5
17. Deshmukh S, Bhirud S (2014) Analysis of audio descriptor contribution in singer identification process. Int J Emerg Technol Advan Eng 4(2):731–737
18. Deshmukh S, Bhirud S (2014) A novel method to identify audio descriptors, useful in gender identification from North Indian Classical Music Vocal. Int J Comp Sci Inform Technol 5(2):1139–1143
19. Shete S, Deshmukh S (2019) Analysis and comparison of timbral audio descriptors with traditional audio descriptors used in automatic Tabla Bol identification of North Indian Classical Music. In: Proceeding of International Conference on Computational Science and Applications, Algorithms for Intelligent Systems, Pune
20. Deshmukh S, Bajaj D, Bhirud S (2015) Audio descriptive analysis of singer and musical instrument identification in North Indian Classical Music. IJRET: Int J Res Eng Technol 04(06):505–508
21. Chordia P (2005) Segmentation and recognition of Tabla strokes. In: 6th international conference on music information retrieval, London, UK, 11–15 Sept 2005
22. Manoj Kumar PA, Sebastian J, Murthy HA (2015) Musical onset detection on Carnatic percussion. IEEE
23. Bello JP, Daudet L, Abdallah S (2005) A tutorial on onset detection in music signals. IEEE Trans Speech Audio Process 13
24. Kshirsagar PR, Akojwar DG (2015) Classification and prediction of epilepsy using FFBPNN with PSO. In: 2015 IEEE international conference on communication networks (ICCN), Gwalior
25. Tampelini L, Boscarioli C, Peres S, Sampaio S (2011) An application of Elman networks in treatment and prediction of hydrologic time series. Learn Nonlinear Models (L&NLM) J Braz Neural Netw Soc 9(3):148–156
26. Jia W, Zhao D, Shen T, Tang Y, Zhao Y (2014) Study on optimized elman neural network classification algorithm based on PLS and CA. Comput Intell Neurosci 2014

Perception of Similarity and Dissimilarity in Hindustani Classical Music

Junmoni Borgohain, Gouri Karambelkar, Raju Mullick, and Priyadarshi Patnaik

Abstract Musicians trained in the Hindustani classical music tradition are usually able to identify the group to which a specific raga belongs. While such classifications make sense in the context of formal musical constructs, it is also essential to examine if there is a perceptual root to such classification. Two studies were conducted under this head. In the first study, musically untrained listeners were asked to group twelve music clips rendered in alaap and madhya laya belonging to two thaats/ragangs and provide descriptions of similarities between the clips in each group. In the second study, an interview was conducted with trained musicians in the Hindustani Classical music tradition to examine the classification parameters used by them. The findings of the first study revealed that untrained listeners were not able to classify the ragas correctly. Similarity judgments between clips were motivated by surface features, whereas trained musicians based their classification on features do not directly fit into the categories of deep features as theorized in Western music. Results suggest that Indian lay listeners use many similar surface features like Western lay listeners, and Indian classification traditions seem to embed deep musical and cultural components unavailable to lay listeners.

Keywords Hindustani classical music · Ragas · Similarity

J. Borgohain (✉) · P. Patnaik
Department of Humanities and Social Sciences, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

P. Patnaik
e-mail: bapi@hss.iitkgp.ac.in

G. Karambelkar · R. Mullick
Advanced Technology Development Centre, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

P. Patnaik
Rekhi Centre of Excellence for the Science of Happiness, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal 721302, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_17

1 Introduction

Musical existence is in a continuum marked by repetition and variation. There are similarities and differences grounded in music that help the listener to distinguish various repetitions, motifs, and thematic variations within it [1]. In this case, musical cues help a listener to successfully categorize and find the relationships between musical motifs and themes [2]. These musical cues play the role of form-bearing dimensions that help the listener to coherently perceive the musical structure by entering in the long-term memory. Furthermore, they are linked to elementary attributes, such as pitch, duration, loudness, and timbre [3].

Consequently, it is believed that comprehending similarities in musical structure requires musical experience and familiarity [4]. Differences exist in defining the relationship between similarity and categorization. Some think that similarity is the foundation on which categorization is built [1, 5], while others contend that they are contrasting concepts [6]. We take the former standpoint in this study, considering similarity as the basis of classification of categories. Categorization behavior depends on the similarity between the item to be categorized and the categories' representations [7]. In other words, when two entities share related properties, they are said to be more similar than the others [1, 2]. Similarity is, therefore, context-dependent, because similarity judgements of entities have a frame of reference; they are relative and intrinsic to comparisons [8].

Musical features, classified as deep and surface features, play an important role in classification tasks. Untrained listeners are based their similarity judgements on "surface features, such as dynamics, articulation, texture, and contour, as opposed to deep features, such as motivic or harmonic relationships" [8]. Other surface similarities related to tempo, rhythmic and melodic texture, pitch register, melodic contour, articulation, and mood were also identified in another study [4]. It is claimed that similarity applies to surface features of the music, whereas categorization takes into account deeper structural elements [6]. However, this understanding of similarity and categorization is limited to the Western classical musical tradition.

In the Hindustani classical music (HCM) tradition, raga is considered as the core framework on which improvisation and composition are performed. Most importantly, there is a wide range of ragas, and they have attracted theorists to classify them under specific classification systems. Classification or grouping of ragas was done on the basis of a set of guidelines indicated in the next section (Sect. 2). Two of the most important classification systems still in practice and used for musical training are thaat and ragang.

2 Thaat and Ragang Classification in Hindustani Classical Music

There are twelve notes in HCM; out of them, seven are shuddha or natural notes (Sa, Re, Ga, Ma, Pa, Dha, and Ni) and five vikrit or distorted notes (re, ga, dha, ni komal, and Ma teevra). When a selection is being made from these notes, it leads to the formation of a thaat. Therefore, a raga is derived from a parent heptatonic scale known as thaats. However, a raga does not need to use every note in a thaat. The ten thaats classified by Pt. Bhatkhande are Bilaval, Khamaj, Kafi, Asavari, Bhairav, Bhairavi, Kalyan, Marwa, Todi, and Poorvi [9]. The thaat classification system is still regarded as one of the vital classification systems for ordering and grouping a wide array of ragas.

On the contrary, Pt. N. M. Khare, a contemporary of Pt. Bhatkhande, observed that many newly evolved ragas could not fit into the mela/thaat classificatory system and introduced a rigorous framework known as the ragang paddhati. Ragas were classified into 30 ragangs based on identifiable note clusters organically known as phrases of notes. The unique pattern of notes (swaras) that are easily recognized is called ragang. The uniqueness of a group of notes in ragang classification depends on the way of using notes, tone accent, vibration, pause, slur, tie, rest, turn, the difference in microtones of notes used, and other complexities.

The brief overview of the two-classification systems suggests that ragang, along with thaat, is essential classification system to make sense of and classify a wide array of ragas that have been influential in the dissemination of music. Thus, it has relevance not only from the point of classification, but also from the points of learner's perspective and pedagogy.

Against this backdrop, this study aimed to understand whether these classification systems are intuitive, i.e., perceptually easy to differentiate, or if they embed deeper non-intuitive features that are only accessible to trained listeners. The first experiment examined if participants correctly categorize ragas belonging to the same thaat and ragang, and what strategies were adopted for such classification. The second study aimed to look at the features used by trained musicians in the HCM tradition to identify and classify ragas.

3 Method

3.1 Study 1

Participants

The participants comprised seventy-seven undergraduate students enrolled in the Bachelor of Technology course at the Indian Institute of Technology Kharagpur. Their age ranged from 16 to 19 years ($M = 17$, SD $= 2.3$). They reported no musical training in their earlier years.

Table 1 Details of stimuli used

Sl. No.	*Ragas*	Notes	Thaat/ragang	Associated time
1.	Kalingada	SrGmPdNS	Bhairav	Morning
2.	Gauri	SrGmPdNS	Bhairav/Gauri	Evening
3.	Bhairav	SrGmPdNS	Bhairav	Morning
4.	Hameer	SRGmMPDNS	Kalyan	Late evening
5.	Chhayanat	SRGmMPDnNS	Kalyan	Late evening
6.	Kamod	SRGmMPDNS	Kalyan	Late evening

Stimuli used

The stimuli consisted of six heptatonic ragas from two thaats and ragang. Kalingada, Gouri, and Bhairav from Bhairav thaat, and Hameer, Chhayanat, and Kamod from Kalyan thaat were chosen. The notes used for the two groups of ragas were the same while their chalans (movements) were different. Care was taken to ensure that they also fit into the same ragang classification, ensuring agreement between two traditions, and greater possibility of identification by lay listeners. For each clip, all the notes of the ragas were covered. Each raga was presented in two distinctive forms, alaap or slow tempo (20 s) and gat or medium tempo (30 s). The clips were vocal renditions generated by a vocalist trained in HCM. The details of the clips are presented in Table 1.

Procedure

The participants in the study completed the music survey using an online link sent to them through email. They were notified earlier and instructed to have headphones available on their systems. A window on 48 h was given to complete the task as a part of their course assignment for which they received course credits. The whole experiment was carried out by designing a computer interface. A click on the link redirected them to a page with elaborate instructions of the experiment. For the classification task, the computer interface had six clips on the top and three boxes below. Participants were asked to listen to the clips as many times as they wanted before putting similar pairs of clips into basket A or B. Clips without any affiliation were dropped in basket C. If participants wanted to change their minds, they could press the key F5 to refresh the task.

Data Analysis

The chi-square method was estimated to examine whether a difference existed while grouping various clips in both tempos. The content analysis method was used to analyze the verbal response of the participants regarding the classification of clips. Categories were formed from the responses. Three judges sat for the codification of the data. The details of the analyses are given in the tables below. For the classification experiment, the hierarchical clustering method was used to analyze how listeners grouped music clips based on similarity.

3.2 Study 2

Participants
For the qualitative interview, sixteen respondents were interviewed. Eight of them were instrumentalists and eight vocalists. Their age ranged from 24 to 31 years, and they had received an average musical training of 20 years.

Procedure
In-depth interviews were conducted with the trained musicians. They were asked what strategies they use to find similarities between two ragas belonging to the same thaat and ragang.
Data Analysis

Content analysis method was used to analyze the responses of the trained musicians.

4 Results and Discussion

4.1 Study 1

The chi-Square results show that the categorization of clips in the *alaap* and *madhya laya* was significant $x^2 = 26.546$, $p < 0.01$. It suggests that the clips were grouped differently from each other in both the tempos. Among the three pairs in the alaap section, "Gauri-Kalingada" received higher responses in pairing as opposed to pairs like "Bhairav-Gauri" and "Bhairav-Kalingada" in the Bhairav thaat. In madhya laya, the opposite was observed: The pair "Gauri-Kalingada" received low ratings compared to "Bhairav-Gauri" and "Bhairav-Kalingada". The clips in the Kalyan thaat, "Chhayanat-Hameer" and "Chhayanat-Kamod", have higher ratings compared to "Hameer-Kamod". However, in the madhya laya, the pairs "Chhayanat-Hameer" and "Chhayanat-Kamod" received lower ratings compared to the alaap. The pair of "Hameer-Kamod" received equal responses in both the tempos. Other pairs in the alaap that received lower ratings were "Chhayanat-Gauri," "Chhayanat-Kalingada," and "Gauri-Kamod". Clips that were rated low in madhya laya were "Gauri-Kamod" and "Hameer-Kalingada" (Table 2).

Using the hierarchical clustering method in the alaap section, three clusters were revealed. The first cluster comprised clips of Bhairav-Gauri, Kalingada-Kamod, Chhayanat-Gauri, and Gauri-Kamod placed in one group. Emotion and melody were the two dominating features identified in this cluster. The second cluster consisted of clips like Gauri-Hameer, Hameer-Kalingada, Bhairav-Kalingada, Hameer-Kamod, and Chhayanat-Kalingada. Emotion, pitch, and melody were highly recognized in the grouping of these clips. Bhairav-Chhayanat, Gauri-Kalingada, Chhayanat-Hameer, Chhayanat-Kamod, Bhairav-Hameer, and Bhairav-Kamod comprised of the third and most distant cluster. This cluster had high responses in emotions identified, pitch, melody, and notes (Fig. 1; Table 3).

Table 2 Chi-square analysis of clips in *alaap* and *madhya laya*

Sl No.	Clips	*Alaap Frequency (%)*	*Madhya laya Frequency (%)*	*Chi-square*
1.	*Bhairav & Chhayanat*	35 (53.0)	31 (47.0)	26.546**
2.	*Bhairav & Gauri*	18 (41.9)	25 (58.1)	
3.	*Bhairav & Hameer*	34 (65.4)	18 (34.6)	
4.	*Bhairav & Kalingada*	18 (42.9)	24 (57.1)	
5.	*Bhairav & Kamod*	29 (56.9)	22 (43.1)	
6.	*Chhayanat & Gauri*	15 (40.5)	22 (59.5)	
7.	*Chhayanat & Hameer*	34 (61.8)	21(38.2)	
8.	*Chhayanat & Kalingada*	15 (39.5)	23 (60.5)	
9.	*Chhayanat & Kamod*	35 (59.3)	24 (40.7)	
10.	*Gauri & Hameer*	24 (52.2)	22 (47.8)	
11.	*Gauri & Kalingada*	37 (60.7)	24 (39.3)	
12.	*Gauri & Kamod*	15 (48.4)	16 (51.6)	
13.	*Hameer & Kalingada*	26 (61.9)	16 (38.1)	
14.	*Hameer & Kamod*	23 (48.9)	24 (51.1)	
15.	*Kalingada & Kamod*	24 (35.3)	44 (64.7)	

**$p < 0.01$

The same method was applied to the clips in the madhya laya, and four clusters were emerged. The first cluster consisted of pairs of Hameer-Kalingada, Hameer-Kamod, and Chhayanat-Hameer. Emotion, melody, and tempo were the overriding characteristics of this group. The second cluster consisted of the three pairs Bhairav-Gauri, Chhayanat-Kamod, and Bhairav-Hameer. Emotion, melody, and rhythm were the highest rated features among the others. Gauri-Hameer, Gauri-Kamod, Chhayanat-Kalingada, Gauri-Kalingada, and Chhayanat-Gauri made the fourth cluster with dominant features like notes, tempo, and melody. The last cluster

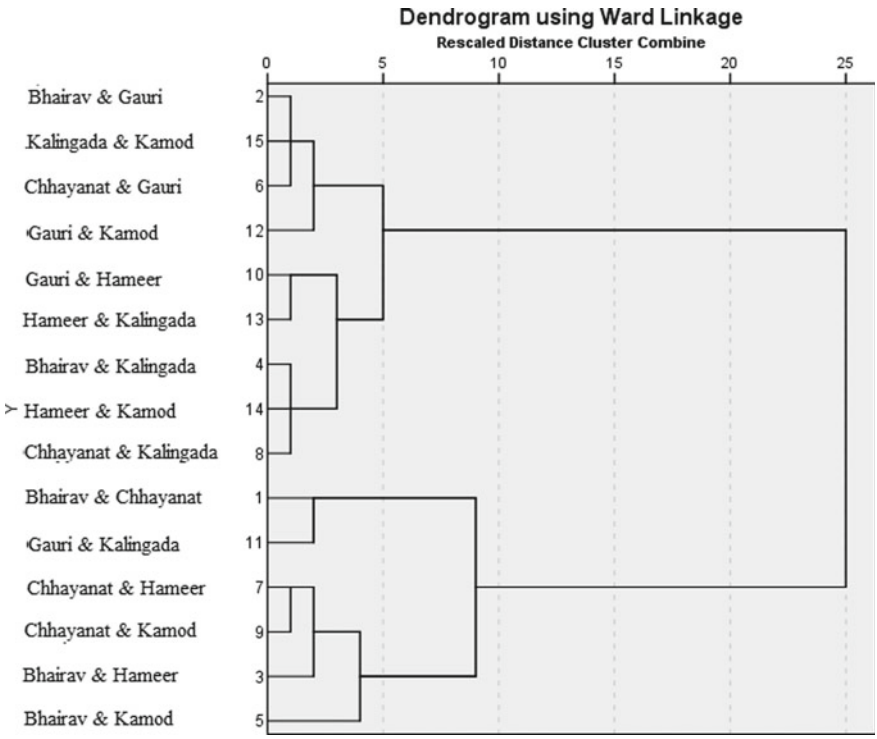


Fig. 1 Hierarchical clustering of clips in *alaap*

Table 3 Qualitative descriptions in the clusters in alaap

Cluster 1	Cluster 2	Cluster 3
Clips 2, 6, 12, 15	Clips 4, 8, 10, 13, 14	Clips 1, 11, 3, 7, 5, 9
Emotion (17), Intensity (4), Loudness (3), Melody (12), Notes (6), Pauses (2), Pitch (8), Rhythm (3), Scale (1), Visual Image (1), Tempo (1)	Emotion (44), Intensity (5), Loudness (2), Melody (11), Notes (10), Pauses (9), Pitch (20), Rhythm (2), Visual Image (4), Tempo (2)	Emotion (79), Intensity (11), Loudness (4), Melody (27), Notes (24), Pauses (12), Pitch (57), Rhythm (4), Scale (6), Visual Image (10), Tempo (2)

had pairs of Bhairav-Kalingada, Kalingada-Kamod, Bhairav-Kamod, and Bhairav-Chhayanat, with emotion, melody, notes, pitch, and tempo as some of the highest rated features (Fig. 2; Table 4).

The hierarchical clustering method revealed that the groups' predicted results were not achieved. Additionally, while grouping the music clips, respondents have predominantly reported emotion as a basis of classification for both tempos. Similarly, other structural objective features, such as intensity, tempo, loudness, and melody, are also identified for categorization of the clips in both tempos. However, these identified features differ both in alaap and madhya laya. Besides emotions, features such as

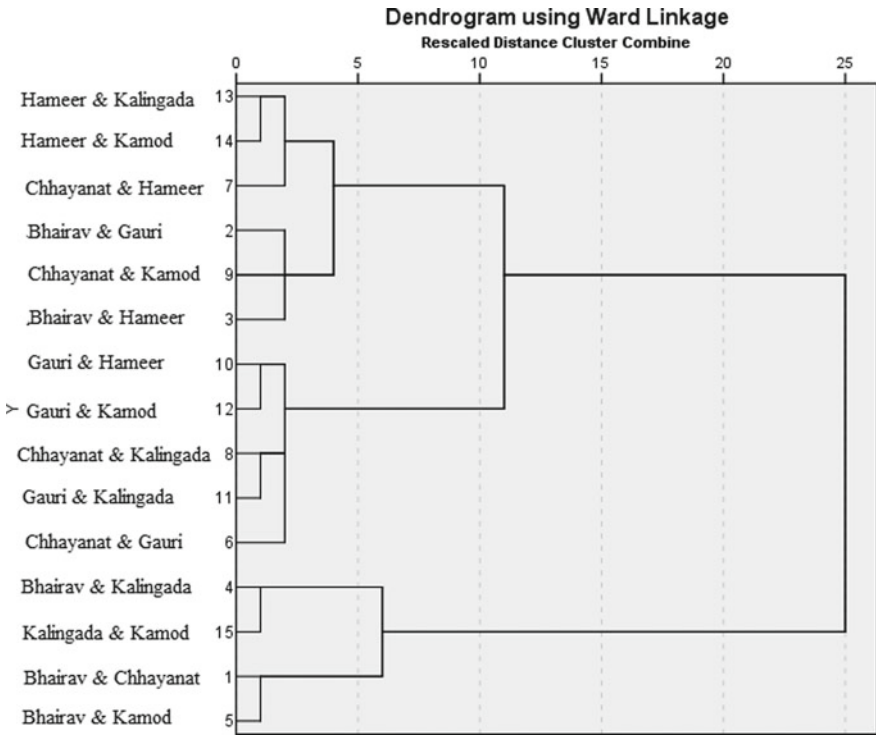


Fig. 2 Hierarchical clustering of clips in madhya laya

Table 4 Qualitative descriptions in the clusters in madhya laya

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Clips 7, 13, 14	Clips 2, 3, 9	Clips 6, 8, 10, 11, 12	Clips 1, 4, 15, 5
Emotion (17), Melody (5), Notes (2), Pitch (1), Tempo (4)	Emotion (14), Melody (9), Notes (7), Pitch (3), Rhythm (9), Scale (2), Visual Image (2), Tempo (4)	Emotion (4), Intensity (1), Melody (7), Notes (11), Pitch (1), Rhythm (3), Visual Image (5), Tempo (10)	Emotion (36), Intensity (1), Melody (16), Notes (14), Pitch (14), Rhythm (7), Scale (1), Visual Image (4), Tempo (10)

pitch, melody, and notes took precedence in alaap, whereas in madhya laya tempo is also a distinct parameter along with the others. The parameters intensity, pauses, and scale are missing from the madhya laya classification categories. From a musical perspective, the participants were looking at identifiable musical sequences in both parts of the clips. Hence, it can be observed that the classification made by the naïve participants was primarily based on surface features rather than deep features. These findings are in line with the earlier findings where untrained participants relied mostly on surface features for classifying music clips [9–11].

In a generalized manner, the mood or the range of emotions created by ragas belonging to the same thaat/ragang is presumed to be similar. However, the results show that in spite of the *ragas* belonging to the same *thaat*/*ragang*, participants were not able to correctly classify them into correct groups. It means that there exists a limit or a threshold of emotions which is much wider, owing to which the participants were not able to easily categorize the ragas belonging to the same thaat/ragang. It also implies that traditional classification of HCM is non-intuitive and not based on popular understanding. It can be inferred that they are based on unique codes that are highly evolved and not necessarily accessible to lay listeners.

Here, the music clips were invaried to the structure, but some of them were considered highly perceptually related as opposed to others. It is, therefore, pertinent to look at the context in which these categorizations have been created [12, 13]. This allows us to attend to the various features involved in perception. The theoretical notion of imprint formation and subsequent empirical observations demonstrate that the abstracted cues are dependent on the musical material [5]. This is also evident from different choices of surface features made by participants for judging similarity between Schoenberg and Beethoven excerpts [8].

4.2 Study 2

An interview was conducted with trained musicians asking them about strategies they would use to classify and categorize ragas with the same thaat and ragang. So, within the musical framework, we find a list of nuanced features adopted by trained virtuosos in the HCM tradition. The trained musicians suggested many features where they would emphasize while categorizing ragas. These themes are classified into ten categories, such as alankaras, emotion, Gharana, movement, notes, octave, phrase rules, tetrachords, and tune. This list of features does not match with the lay people's perspective of categorizing the various clips. However, surface features were also indicated by the trained musicians (Table 5).

Clearly, prioritization of feature categories by both the groups can be observed where the latter group focusses on the features driven by a top-down approach as opposed to untrained participants. This leads us to derive at the premise that the classification of HCM is non-intuitive. The framework through which trained musicians see *ragas* might be socially constructed rather than naturally evolved.

5 Conclusion and Future Work

This study gave us an overview of the features used by trained and untrained participants to classify and distinguish between ragas. Untrained participants mostly use surface features like emotions, melody, notes, pitch, tempo, and rhythm among many others to classify ragas. On the other hand, trained participants use categories like

Table 5 Classification categories used by trained Hindustani classical musicians

Themes	Features
1. *Alankaras* (Ornamentation)	– Vadi-samvadi (sonant) – Andolan – Theheraw
2. Emotion	– Emotion evoked
3. *Gharana* specification	– Same raga rendered in two different gharana styles
4. Movement	– Chalan, detailed study of various bandishes in those ragas to understand the nuances – Gati/pace of the raga will be different, in case of homogeneous ragas – Look at chalan minutely-use of notes in aroha-avaroha (ascendance-descendance of notes)
5. Notes	– Study of a raga family at a time – Learn various vocal bandishes – Swara lagaw – Use of notes in a raga (komal ni in ramkali) – Avirbhav-tirobhav (Avirbhav is straightforward display of the core of the raga, whereas tirobhdva is a phrase or set of phrases that completely distort the picture of the raga) – Nyas (rest notes) – Mukhyang (essential notes) – Swara sangati
6. Octave	– Poorvang/uttarang – Bandish—dos and don’ts while rendering
7. Phrases	– Important phrases of the family and the particular raga – Phrases (graha-ansha-nyas included in phrase) – Sequence of phrases—grammar of the raga – Sangati—importance of notes – In case of ang/chalan raga—which ang is prominent – In case of jod raga (combination of two ragas) proportion of the two ragas, then look whether that raga has a peculiar phrase as its own—in ragas like Nat or Kanada types it is more compacted—it is more difficult in Uttarang – Very close ragas—see root of the raga—limited phrases used – There are some phrases in the ragas which are neither "ang" nor a phrase of that raga—we need to identify them
8. Rules	– Don’ts of the raga
9. Tetrachords	– Vadi-samvadi
10. Tune	– Listening to the tune – Listening to a raga as a dhun

rules, notes, tetrachords, alankaras, and phrases to classify ragas which do not fit specifically into the deep features theorized in Western music. These findings are in line with the classification features, “surface” features used in Western classical music. The study has its limitations since its scope does not permit an enquiry into

the surface/deep features of Indian classical music and how these relate to Western classical music, an exploration that needs to be carried out in the future.

References

1. Cambouropoulos E (2009) How similar is similar? Musicae Sci 13(1_suppl):7–24
2. Cambouropoulos E (2001) Melodic cue abstraction, similarity, and category formation: a formal model. Music Percept: An Interdiscip J 18(3):347–370
3. McAdams S (1989) Psychological constraints on form-bearing dimensions in music. Contemp Music Rev 4(1):181–198
4. McAdams S, Vieillard S, Houix O, Reynolds R (2004) Perception of musical similarity among contemporary thematic materials in two instrumentations. Music Percept: An Interdiscip J 22(2):207–237
5. Deliège I (2001) Introduction: similarity perception-categorization-cue abstraction. Music Percept 18(3):233–243
6. Ziv N, Eitan Z (2007) Themes as prototypes: similarity judgments and categorization tasks in musical contexts. Musicae Sci 11(1_suppl):99–33
7. Goldstone RL (1994) The role of similarity in categorization: providing a groundwork. Cognition 52(2):125–157
8. Lamont A, Dibben N (2001) Motivic structure and the perception of similarity. Music Percep: An Interdiscip J 18(3):245–274
9. Ranade A (2006) Music contexts: a concise dictionary of Hindustani music. Promila and Co. Publishers in association with Bibliophile South Asia, New Delhi and Chicago, pp 273–274
10. Deliège I (1989) A perceptual approach to contemporary musical forms. Contemp Music Rev 4:213–230
11. McAdams S, Matzkin D (2003) The roots of musical variation in perceptual similarity and invariance. In: Peretz I, Zatorre R (eds) The cognitive neuroscience of music. Oxford University Press, Oxford, pp. 79–94
12. Goldstone RL, Medin DL, Halberstadt J (1997) Similarity in context. Mem Cognition 25(2):237–255
13. Rips LJ (1989) Similarity, typicality and categorization. In Vosniadou S, Ortony A (eds) Similarity and analogical reasoning. Cambridge University Press, Cambridge, pp. 21–59

Analytical Comparison of Classification Models for Raga Identification in Carnatic Classical Audio

Sathwik Acharya, Vihar Devalla, Om Amitesh, and Ashwini

Abstract There is this famous quote which goes as follows "MUSIC is the divine way of portraying the most beautiful things about this world." With that being said, the diversity in this language of music is immense, to say the least. Broadly, one would be well aware of the classification between Indian classical music and western music. In music information retrieval (MIR), raga classification has a tremendous part role in understanding the fundamentals of Indian classical music and in a multitude of other tasks like database organization of music files to music recommendation systems. The paper encompasses a variety of techniques like ANN, LSTM, and XGBoost models for the task of raga identification. The work is initially carried out on a set of 10 ragas and then extended to 20 ragas. Both the tasks showed impressive results with an accuracy of 99.56% and 99.43% for a set of ten and twenty ragas, respectively. The process was carried out on the ragas pertaining to Carnatic music, a division of Indian classical music. The data samples for the same were obtained from a standard data set.

Keywords LSTM · ANN · XGBoost · Raga identification · MIR

S. Acharya (✉) · V. Devalla · O. Amitesh · Ashwini
PES University, Bengaluru, Karnataka 560085, India

V. Devalla
e-mail: PES2UG19CS453@pesu.pes.edu

Ashwini
e-mail: ashwinib@pes.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_18

1 Introduction

1.1 A Glimpse into Indian Classical Music and Its Technical Aspects

When one delves into Indian classical music, the one stark thing that stands out is the existence of ragas. In simple words, ragas are a collection of features (like aroha-avaroha) that can uniquely identify a given song. It can be broadly classified into Hindustani music (North Indian music) and Carnatic music (South Indian music). It is very interesting to note that the two systems show more common features than differences.

The arrangement of notes of a raga in ascending order is called the arohana of that raga and the arrangement of notes of a raga in descending order is called the avarohana. Different notes are called swaras in Indian classical music. The arohana and the avarohana form the foundation in which the rest of the song is composed. One must also remark the fact that not a lot of attention is given to western musical concepts like chord recognition, etc., in Indian Classical music.

In contrast to western music, raga is analogous to melody but has more intricate features than that. It is possible for two ragas to be similar but the musical effects they produce show a strong distinctive feature. This can be drawn to several reasons like the temporal arrangement, gamakas (arrangement of the notes in an oscillatory motion) or the pakad (phrases of swaras that are unique to a raga).

It is strongly seen that the task of raga classification is linked to sequential tasks. As a result, the models tested are LSTM, XGBoost, and feedforward neural network (ANN). As you can see, we have compared machine learning algorithms with deep learning models. The advent of deep learning models has changed the entire landscape by providing state-of-the-art results in various domains. However, it is important to note that XGBoost, an ML algorithm provides state of the results here (over 99%) and captures better features in contrast to DL models.

2 Related Work

A variety of methods has been implemented in raga classification using machine learning algorithms or deep learning approaches or a combination of both. The earliest work on this can be traced to [1]. It is the first raga identification model that uses hidden Markov models (HMM) with string matching algorithm approach to classify songs according to their raga. This was followed by the method of pitch class distributions and harmonic class profiles together for classification using support vector machines (SVM) as in [2]. However, a glaring problem with this is that it discards temporal information (for instance, musical timing).

This was followed by an approach involving convolutional neural networks [3] to classify raga. Shetty and Achary [4] makes use of extracting the arohana and

avarohana patterns to classify music into its respective raga. The drawback of this approach is that, despite arohana and avarohana form the base on which the entire raga is structured, they cannot alone capture the richness unique to a raga. Gulati et al. [5] makes use of time-delayed melody surface (TDMS) which addresses the above problems. This model has shown state-of-the-art performance of about 98% accuracy on a dataset of 30 ragas.

There is also [6], which is a comparative study of machine learning different approaches to the problem, much like this paper, implementing the use of C4.5, Bayesian, random forest, and K^* classifiers. Lastly, one of the recent papers [7] uses the LSTM-RNN approach to classify Indian classical music. It uses the LSTM approach for two specific tasks, sequence classification and sequence ranking and retrieval (to rank similar sequences and thereby fine-tune the classifier). This model has achieved an impressive accuracy of 97% on a dataset of 10 ragas.

3 Methodology

3.1 Preprocessing

It is quite often for a musical arrangement to have an ensemble of different elements come together. These might range from the percussion kits to the instruments and the vocals. For the task of identification of raga, two approaches have been tried—one without separation of voice from the accompaniment and the other with the separation of vocal from the accompaniment. With that in mind, Spleeter is used to isolate the vocal from the accompaniment part. It is a neural network-based source separation library with pretrained models that makes it easy to train source separation models and provide state-of-the-art models for performing a plethora of source separation operations.

3.2 Feature Extraction

There are a multitude of features that come into picture for the analysis of audio samples. Here, a popular Python library called Librosa is used for the same. It includes the prerequisite packages to build a functional and operational music information retrieval (MIR) system. The files in our dataset were of the standard .wav format.

Visual representation of audio samples is typically done via the help of spectrograms. The spectral features are as follows:

- Spectral roll-off
- Spectral bandwidth
- Chroma feature

Table 1 Tabulated important features along with their range

Feature	Definition	Equation	Range of the values
Spectral centroid	It indicates the frequency at which the energy of the spectrum is focused upon	$\frac{\sum_k S(k) f(k)}{\sum_k S(k)}$ where $S(k)$ is the spectral magnitude at frequency bin k, $f(k)$ is the frequency at bin k	284.59–7527.58
Zero crossing rate	It is the rate at which the signal changes its sign, that is, the rate at which the value of our signal changes from negative value to 0 or from 0 to positive value and so on	$\frac{1}{T-1} \sum_{t=1}^{T-1} II\{s_t s_{t-1} < 00\}$ s_t is the signal of length t II$\{X\}$ is the indicator function(=1 if X is true, else 0)	0.005–0.7203
MFCC	The mel-frequency cepstral coefficients (MFCCs) of a waveform are a small set of features (usually about 10–20) which elegantly describe the overall shape of a spectral waveform	$\sum_{u=0}^{N} X[h\tau + n] W[n] \mathrm{e}^{-j2\pi kn/N}$	−58.36 to 13.261 (for MFCC7)

The pitch classes of chroma features in Carnatic music are analogous to the swara notes (Sa, Ri, Ga, etc.). In comparison to Western music, this is equivalent to C, C#, D, and so on. The other features are tabulated in Table 1.

The reason for choosing the above features can be traced to the fact that they capture most, if not all the information about a given audio file.

4 Analysis of the Classification Models Used

It is important to note that the approach for the process of feature extraction involves taking the features from the audio file by considering the 10 different samples of 30 s duration from the song, i.e., three samples from the beginning, four from the middle, and three from the end parts of the song, respectively.

The above extracted features are extracted and fed into a CSV file. Following are the different models which we have used for the task.

4.1 Artificial Neural Network (ANN)

A feedforward neural network architecture which is used for most supervised learning tasks is chosen.

The model consists of four fully connected layers, in entirety, with two hidden layers. The model begins with an input layer, which takes a batch of rows (each row consisting of 27 values as per the features of the clip) from the csv file as input and gives an output vector of size 256. This is then passed into the first and subsequently, the second hidden layers of size 128 and 64 neurons. The input and the two hidden layers implement a rectified linear unit (ReLU) activation. The final output layer with a softmax activation gives us a vector of size 10 (for each clip input), containing the probabilities of the 10 ragas the input row can belong to.

- **Model Training and Optimization**

The model is trained for 500 epochs, with a batch size of 128 (128 rows or clips of songs as input in one training step) with a learning rate of 0.001.

- **Loss Function**

The loss function implemented for this method is the categorical cross-entropy (sparse variant in this case, as we encoded our label ragas as integers). It is apt to use this loss function as this is a classification problem with more than two classes. The categorical cross-entropy loss is given as:

$$\text{Categorical cross-entropy} = \sum_{i=1}^{N} y_{i,j} \cdot \log(p_i)$$

Here, N corresponds to the number of classes (number of ragas in our case), and $y_{i,j}$ is the indicator function that assumes 1 when i and j are equal and j is the training label (or the target raga in this case) for the given data from the .csv file, p_i corresponds to the ith entry of the output vector from our model which corresponds to the probability of the select clip to be of that particular raga.

- **Optimizer**

Adam optimizer, an optimizer popular in training models for classification purposes is used. It is preset with the values for beta 1 and beta 2 as 0.9 and 0.999, respectively, and the value for epsilon being 1e−07 (Fig. 1).

4.2 Long Short-Term Memory(LSTM)

Among the many models used, one of them was long short-term memory (LSTM). Long short-term memory model since its inception is used mainly for sequence

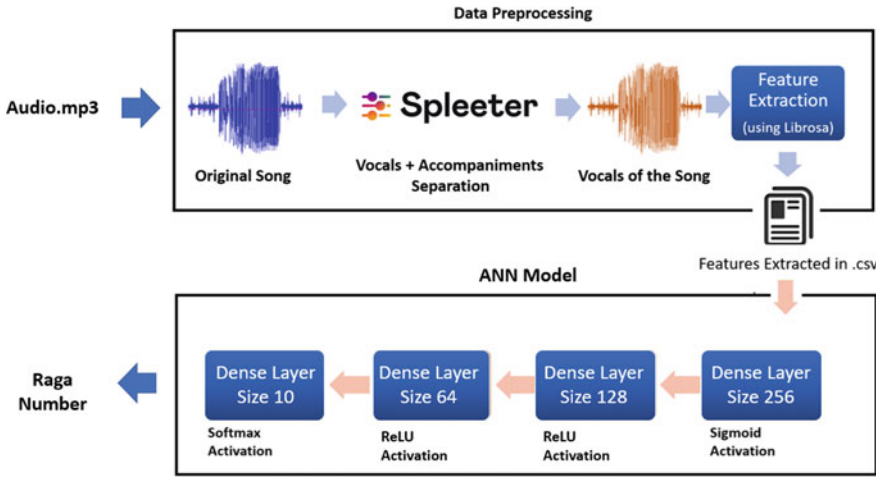


Fig. 1 A flowchart depicting the ANN (feedforward) model including the data preprocessing steps

prediction problems and fits very well with the input samples used in the project. It relies on its feedback connections and ability to learn long-term dependencies. After trying various configurations to get the best results, we settled with a four-layered neural network having 2 LSTM layers followed by a flattening layer and a dense layer having 10 or 20 units corresponding to the number of ragas required to be classified. The model was sent into training for about 300 epochs with a batch size of 35 and an initial learning rate of 0.001. The size of the training dataset was 1680 samples and a test set of 560 samples (Fig. 2).

The first layer using LSTM with a ReLU activation function converts the data into a 128-dimensional vector and a dropout regularizer to probabilistically exclude the input and recurrent connections in the network. This solves the predominant problem (common in these tasks) of overfitting and improving model performance. The second layer has a similar working except for the fact that it converts the input it gets to a 64-dimensional vector and sends it to a flattening layer. Here, data for each sample is converted into a 1D array and sent to the dense layer to get a single prediction value to each raga for the given sample. The loss function used is the

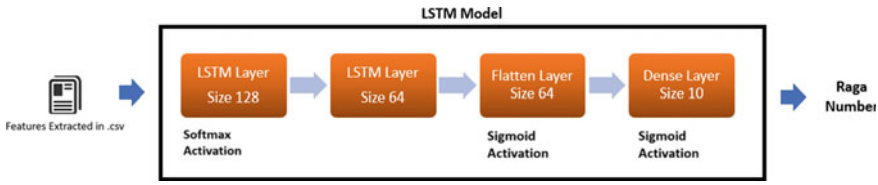


Fig. 2 Architecture of LSTM network

same as used in the ANN model (sparse categorical cross-entropy) and the Adam optimizer to better the results of the model.

4.3 XGBoost

Based on the popular decision tree-based ensemble algorithm that uses a gradient boosting framework, XGBoost tends to outperform all other models when it comes to small-sized structured data. In prediction scenarios, however, involving unstructured data, neural networks (i.e., artificial neural networks ANN) shows state-of-the-art results compared to all other algorithms. It (XGBoost) is an ideal blend of software and hardware fine-tuning techniques to yield impressive results using less computational resources. Another important thing to remark is its fast and efficient computational speed. The place where XGBoost stands out can be traced to its high performing GBM framework. This is realized via system optimization and enhancement of the algorithms. XGBoost uses DMatrices in place of general forms of data like numpy arrays or Pandas Dataframes. These DMatrices can contain both the features and the targets. The implementation of XGBoost carries with it a number of hyperparameters. Tuning these hyperparameters depends on the task at hand and can yield better results. It is important to note that hyperparameters often take on a set of values, with glaring exceptions being things like regularization values. Therefore, in order to avoid the intricacies involved in hybrid discrete optimization, it is a common practise to tune the hyperparameters by discretizing the values of all hyperparameters in question. For instance, five hyperparameters will be fine-tuned for the XGBoost model. The extent of possible values that will be taken into consideration can be seen in Table 3.

In very simple terms, grid search (GS) can be traced to a brute force approach that looks for all the possible combinations of hyperparameters in some discrete order. This is followed by the computation of the cross-validation loss for each combination and finds the optimal values in this manner. The optimal values received can be seen in Table 2. These values were then fed into the model. Figure 3 gives a visual representation of the importance of the features (27 in total) for the XGBoost model (Table 3).

Table 2 Optimal values for hyperparameters

Parameter	Optimal value
eta	0.05
Max depth	3
Gamma	0.0
colsample_bytree	0.7
Min_child_weight	1

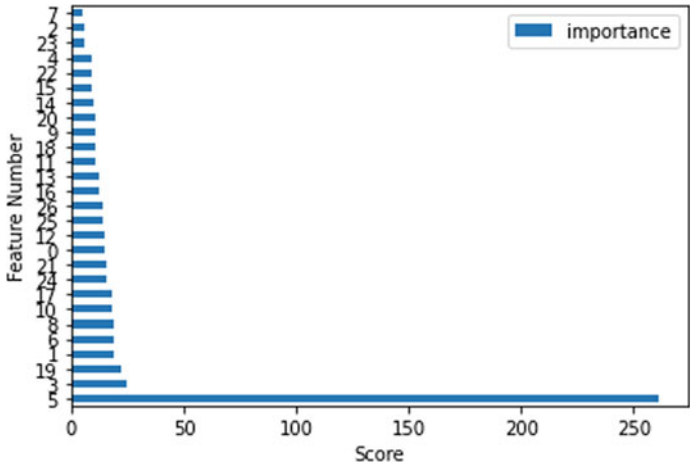


Fig. 3 Graph of feature importance

Table 3 Grid search for XGBoost

Parameter	Property	Range of values
eta	It controls the learning rate	[0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
Max depth	It is the maximum number of nodes allowed from the root to the farthest leaf of a tree	[3, 4, 5, 6, 8, 10, 12, 15]
min_child_weight	It is the minimum weight required in order to create a new node in the tree	[1, 3, 5, 7]
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree	[0.0, 0.1, 0.2, 0.3, 0.4]
colsample_bytree	It corresponds to the fraction of features (the columns) to use	[0.3, 0.4, 0.5, 0.7]

5 Results

We must emphasize on the fact that the entire collection of ragas for our task was extracted from the Dunya Web site [5, 8]. Only those ragas were chosen which had around 15 songs. This was done to ensure homogeneity of the dataset. The list of the ragas can be seen in Table 6.

The following approaches were employed for the raga classification task. The results can be divided into two main branches:

1. Without separating the vocals from the accompaniment: This approach involves dividing the song into different samples and extracting the features for each of these segments. The results for the different models can be seen in Tables 4 and 5.

Table 4 LSTM model, ANN model

LSTM model				
Approach	Validation loss (10 ragas)	Validation accuracy (10 ragas) (%)	Validation loss (20 ragas)	Validation accuracy (20 ragas) (%)
A	0.1197	96.00	0.2309	94.17
B	0.2162	93.07	0.4123	89.17
ANN model				
Approach	Validation loss (10 ragas)	Validation accuracy (10 ragas) (%)	Validation loss (20 ragas)	Validation accuracy (20 ragas) (%)
A	0.0402	98.67	0.1641	96.50
B	0.1311	97.07	0.1769	95.52

Table 5 XGBoost model

Approach	Validation accuracy (10 ragas) (%)	Validation accuracy (20 ragas) (%)
A	99.56	99.43
B	99.49	99.21

2. By separating the vocal from the accompaniment: This approach again involves dividing the song into different samples and extracting the features for each of these segments. The results for the different models can be seen in Tables 4, 5 and 6.

The images below give us a visual representation of the accuracy and losses associated with the ANN and LSTM model, respectively (Figs. 4, 5, 6 and 7).

Table 6 List of ragas used

Name of raga
Abhogi Darbar
Ananda Bhairavi Varali
Atana Jaunpuri
Begada Vasanta
Behag Yamuna Kalyani
Bhairavi Jaganmohini
Bilahari Shanmukha Priya
Hamsadhwani Yadukula Kambhoji
Hindolam Todi
Purvikalyani Mayamalava Gaula

Note: A—Refers to the approach where no vocal separation is done from the accompaniment
B—Refers to the approach where vocal separation is done from the accompaniment

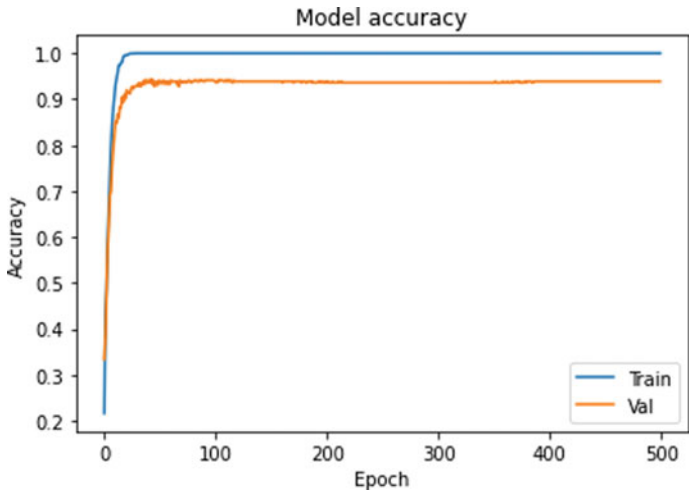


Fig. 4 Accuracy plot for ANN

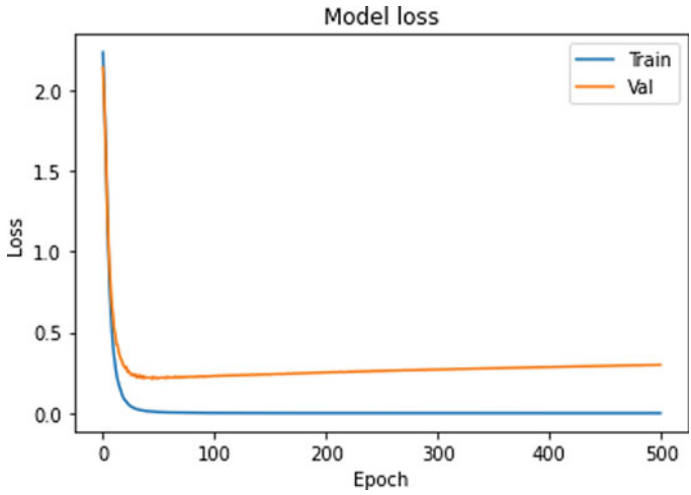


Fig. 5 Loss plot for ANN

6 Conclusion and Future Scope

Drawing inferences from the table and the work, we can see that the first approach which is without separating the voice from the accompaniment provides better results. This can be attributed to the fact that the accompaniment part can add significantly to the features of a given raga which is an unusual observation. Among these approaches,

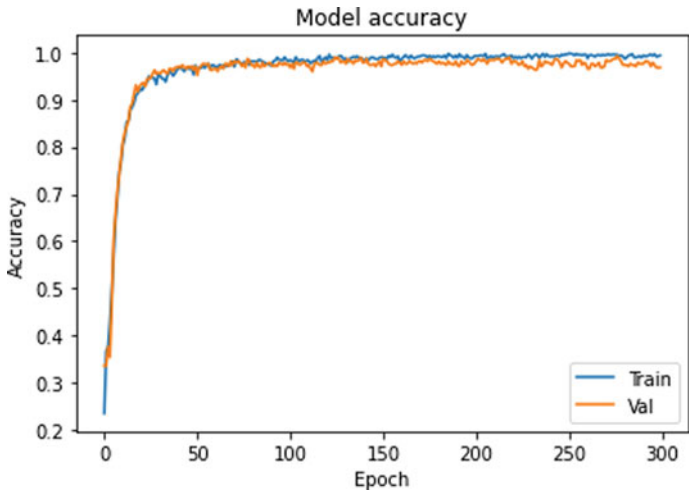


Fig. 6 Accuracy plot for LSTM

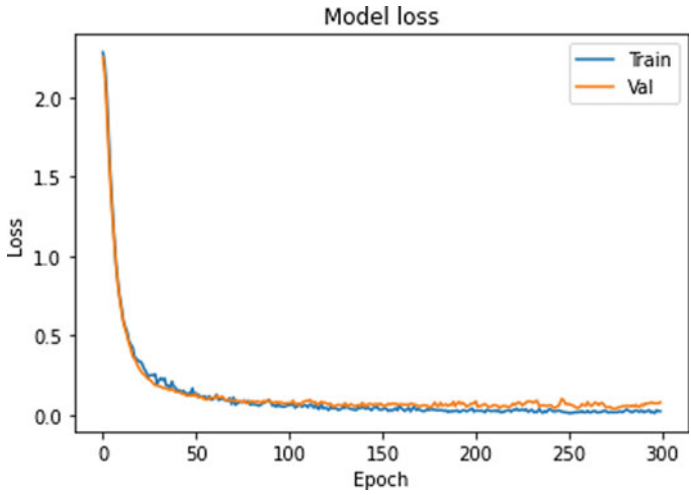


Fig. 7 Loss plot for LSTM

the XGBoost model provided the best results. This can be attributed to its effectiveness of tree pruning, cross-validation, and regularization. The same approach can be extended to identify more number of ragas from the Hindustani Classical Genre as well as from instrumental audio.

References

1. Pandey G, Mishra C, Ipe P (2003) Tansen: a system for automatic raga identification. In: Proceedings of Indian international conference on artificial intelligence, pp 1350–1363
2. Chordia P, Rae A (2007) Raag recognition using Pitch-class and pitch-class dyad distributions. In: Proceedings of ISMIR, 2007, pp 431–436
3. Madhusdhan ST, Chowdhary G (2018) Tonic independent Raag classification in Indian classical music
4. Shetty S, Achary K (2009) Raga mining of Indian music by extracting arohana-avarohanapattern. Int J Recent Trends Eng 1(1)
5. Gulati S, Serrà J, Ganguli KK, Sentürk S, Serra X (2016) Time-delayed melody surfaces for raga recognition. In: Proceedings of the 17th international society for music information retrieval conference (ISMIR), New York, USA, pp 751–757
6. Sharma H, Bali RS (2015) Comparison of ML classifiers for Raga recognition. Int J Sci Res Publ 5
7. Madhusudhan ST, Chowdhary G (2019) DeepSRGM—sequence classification and ranking in Indian classical music via deep learning. In: Proceedings of the 20th international society for music information retrieval conference. ISMIR, Delft, pp 533–540
8. Gulati S, Serrà J, Ishwar V, Sentürk S, Serra X (2016) Phrase-based rĀga recognition using vector space modeling. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), Shanghai, pp 66–70. https://doi.org/10.1109/icassp.2016.7471638

Multimodal Sentiment Analysis of Rabindra Sangeet Through Machine Learning Techniques

Sandipan Ganguly, Dipankar Das, Akash Modak, and Samanway Chakraborty

Abstract The voluminous works of poet Rabindranath Tagore reflect a wide VIBGYOR of his creative persona. Within this great repertoire of creation, the songs he wrote and as well as rendered the tune have a significant place. The entire corpus of these songs which is better known as Rabindra Sangeet has a number of motifs which are categorized by the term 'parjaay' or class by the poet himself. Each of this class reflects a certain mood of human psyche. In the present work, machine learning techniques were used to analyse the songs belonging to four of the major class or 'parjaay' of the entire corpus of Rabindra Sangeet. Both lyrics and audio were separately used as the dataset, and related features were extracted to be fed into some of the existing machine learning algorithms. The results of these analysis were compared. We propose that this hitherto unknown means of analysing the works of Tagore will open new avenues to render further deep level analysis of the inherent semantics of his works.

Keywords Sentiment analysis · Rabindra sangeet · Multimodal · Deep learning

1 Introduction

Seldom is the case that the literary work of an individual embodies the deep-rooted philosophical vision of an entire nation. Poet Rabindranath Tagore and the whole plethora of his creation really signifies such a trait. Among the palette of his creative works, the songs have special significance. The poet himself categorized them into a number of class or 'parjaay' as it is better known, each belonging to a specific thematic motif. The computational analysis of the lyrics and the tune may establish traits to find out inherent semantic and pragmatic features embedded within. In recent years, similar sentimental analysis of natural language text has become an important research topic to analyse the complex semantic features embedded within classical

S. Ganguly · A. Modak (✉) · S. Chakraborty
Department of Computer Applications, Heritage Institute of Technology, Kolkata, India

D. Das
Department of Computer Science, Jadavpur University, Kolkata, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_19

literary texts. In this work, Tagore songs were used as the dataset as they are widely considered as sort of a gold standard of musical verses. This is a primary work that involves analysis of Tagore songs where lyrics and the musical files (waveforms) taken separately for machine learning analysis. We used five models, namely naive Bayes classification, support vector machine (SVM), single-layer perceptron and multi-layer perceptron and recurrent neural network (RNN) to analyse the lyrics. For musical files, however, we used two suitable models, namely RNN and LSTM. Using proper training and test classes, the identification of individual 'parjaay' was tested. For dearth of text features from the lyrics dataset, we also projected certain visualizations depicting as: (i) the songs with most word count and (ii) most frequent words pertaining to each 'parjaay'.

The paper is organized through the following sections. Section 2 reviews some related works on musical sentiment analysis. Section 3 describes the multimodal data or corpus used for the present work. The detailed preprocessing, feature extraction and results obtained after fitting the models are discussed in Sects. 4 and 5. Finally, Sect. 6 draws the conclusion of the present work and discusses certain avenues of future scope of the work.

2 Related Work

The present work draws inspiration from some of recent research works that attempted to bridge between the semantic motifs embedded within literary works and the machine learning models to mathematically project the inherent semantics. Koduri and Indurkhya [1] researched to classify the moods inherent in South Indian classical music, i.e. carnatic music. The primary objective of their experiment was to verify the raagas that really depict a dominant emotion. Some works on mood classification inherent to western music have been performed using audio signals and lyrics [2–4]. Patra, Das and Bandyopadhyay experimented to obtain a semi-supervized learning paradigm to classify the mood inherent in songs (2013) [5, 6]. They emphasized the semantics of the Hindi songs (2015). Yang, Lin, Su and Chen adapted a regression approach to recognize music (2008) [7]. Xie, Rastogi and Chang adapted a model to study and analyze word-level and character-level language models for Shakespearean sonnet [8].

3 Dataset

The present work involves multimodal analysis of two different types of datasets. For lyrics, we have used the Rabindra Sangeet lyrics corpus from the Society for Natural Language Technology Research (https://rabindra-rachanabali.nltr.org/). For the audio files, we have used Gitabitan Archive and Shruti Gitabitan DVD repository.

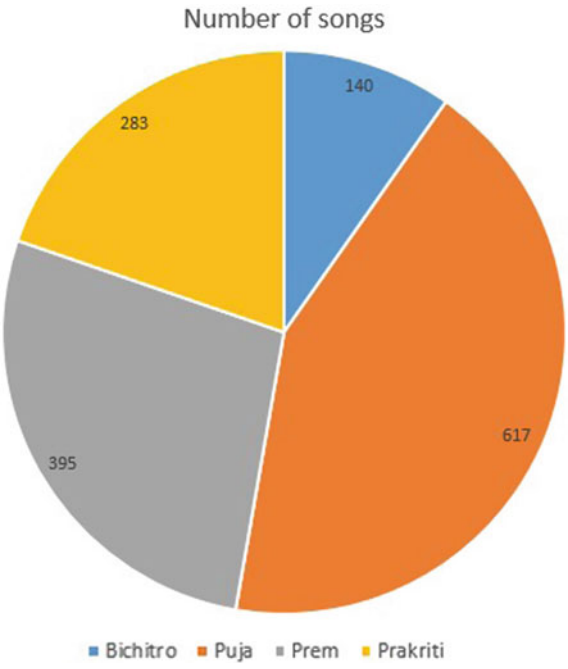


Fig. 1 Pie Chart of the number of songs used for lyrics against various 'parjaay'

The basic idea behind is to recognize the 'parjaay' of the song based on its enumerated class level. We first enumerated the 'parjaay' with proper class levels (Prem-1, Bichitro-2, Puja-3, Prakriti-4). The number of song lyrics used are depicted in Fig. 1.

4 Methodology

4.1 Feature Extraction

The features of both of these two different types of data, namely lyrics and audio, are extracted with suitable means. For the lyrics, we have used nltk and gensim packages. Tf-idf and n-gram features with n value 1 and 2 are used as features. For music data, we have used Librosa package. The extracted features were then fed into a number of deep learning paradigms. We used the Google Colab platform for working with the extracted features both of these two types of data.

4.2 Pre-processing of Data

4.2.1 For Lyrics

Lyrics data was scraped from the mentioned URL along with the 'parjaay' or genre. The preprocessing of this data involves deleted the unwanted characters including ?, /, @, #, $, %, & etc. as these characters will not contribute to the genre detection algorithm. After that the pre-processed lyrics was enumerated to some numerical values.

4.2.2 For Audio

At first, the audio features were extracted using the Librosa library that resulted in generating numerous features. Then, through PCA, the following dominant feature variables were found having the most variations.

1. MFCC
2. Zero-crossing rate
3. Spectral roll-off
4. Spectral flux
5. Pitch
6. Spectral entropy
7. Chroma features.

5 Model and Experimental Results

5.1 Lyrics Data

The varied nature of the data made us test through various forms of machine learning methodology and find the relatively best fit for the type of data. For deployment, Scikit-learn is used for naive Bayes, SVM, single and multi-layer perceptron models, and for RNN, TensorFlow is used as the software paradigm. For the lyrics, we have used five models, namely naive Bayes classification, support vector machine (SVM), single-layer perceptron and multi-layer perceptron and recurrent neural network (RNN) to analyse the lyrics. The following settings are used in the respective models.

1. Naive Bayes classifier:
   (a) Multinomial Naive Bayes

2. Support vector machine:
   (a) Linear support vector machine with stochastic gradient descent classifier
   (b) Learning rate—adaptive
   (c) Loss—hinge
   (d) Validation fraction—0.2
3. Single-layer perceptron:
   (a) Penalty—elasticnet
   (b) Max-iter—5000
4. Multi-layer perceptron:
   (a) Hidden layer—15
   (b) Activation function—identity
   (c) Solver—adam
5. Recurrent neural network:
   (a) Layers—3
   (b) Activation function—relu, output layer - softmax
   (c) Loss—Binary cross-entropy
   (d) Optimizer—Adam optimizer
6. LSTM:
   (a) Layers—3
   (b) Activation function—relu, output layer—softmax
   (c) Loss—Binary cross-entropy
   (d) Optimizer—Adam optimizer
   (e) Epochs—1000
   (f) Dropout
      I. First layer—0.2
      II. Second layer—0.1

The results in terms of F1-score, recall and precision are depicted from Figs. 2, 3, 4, 5 and 6, respectively.

The overall accuracy measure of lyrics data is plotted in Fig. 7 against all the mentioned learning methodologies. The interesting observation through the previous plots reveals that 'Bichitro' parjaay has a significant deviation in terms of all three important learning parameters. Now, this 'parjaay' or class as per the Bengali meaning suggests itself is an amalgamation of multiple sentimental motif. That multiplicity is well caught here through mathematical means.

Numeric measures for precision, recall and F1-score as found by the application of different learning models are depicted from Tables 1, 2, 3, 4 and 5. The low value for 'bichitro' is attributed for the data inadequacy, and it is being an amalgamation of different themes.

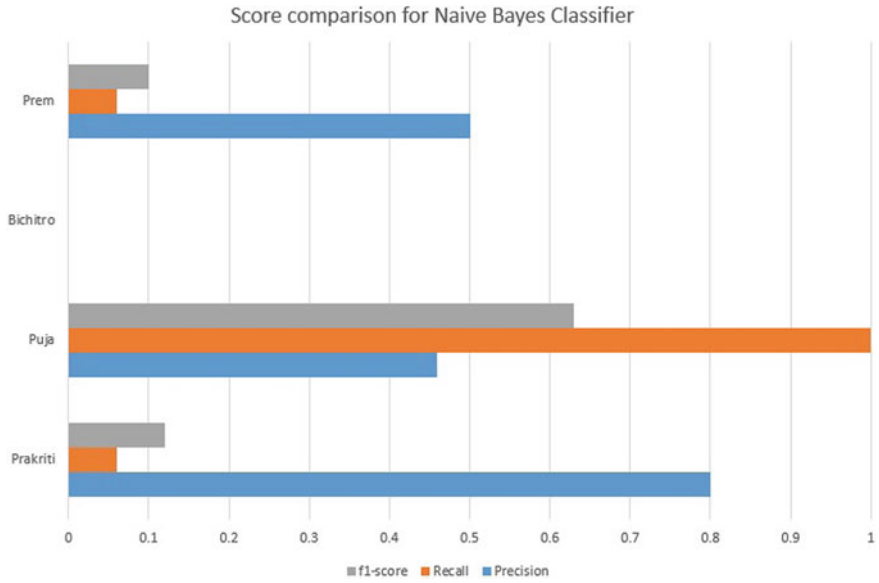


Fig. 2 Comparison of lyrics feature across multiple parjaay through naive Bayes classifier

Fig. 3 Comparison of lyrics feature across multiple parjaay through support vector machine

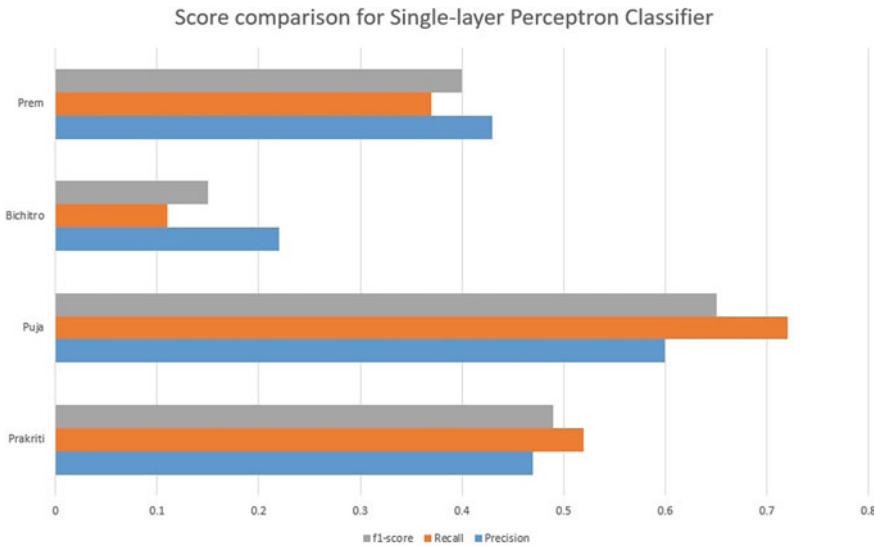


Fig. 4 Comparison of lyrics feature across multiple parjaay through single-layer perceptron

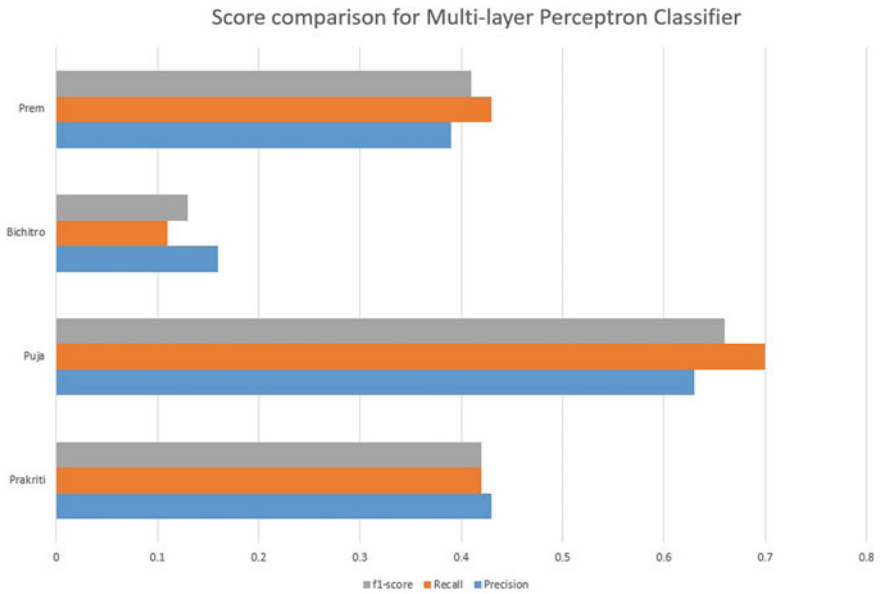


Fig. 5 Comparison of lyrics feature across multiple parjaay through multi-layer perceptron

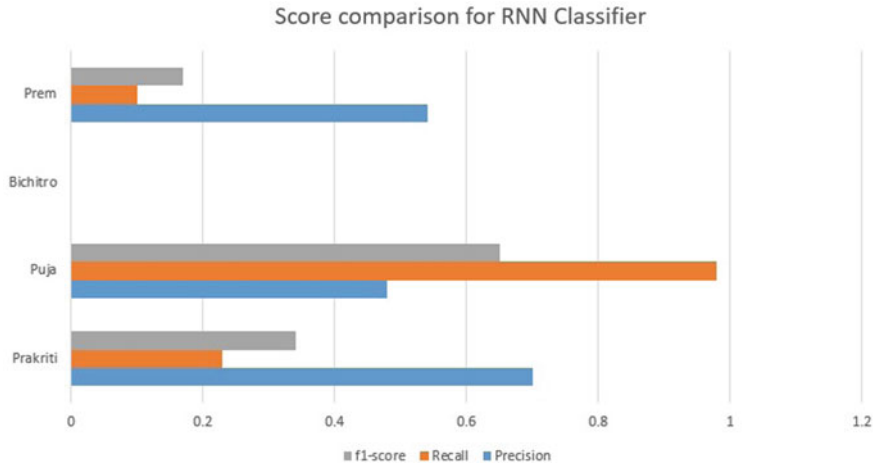


Fig. 6 Comparison of lyrics feature across multiple parjaay through RNN

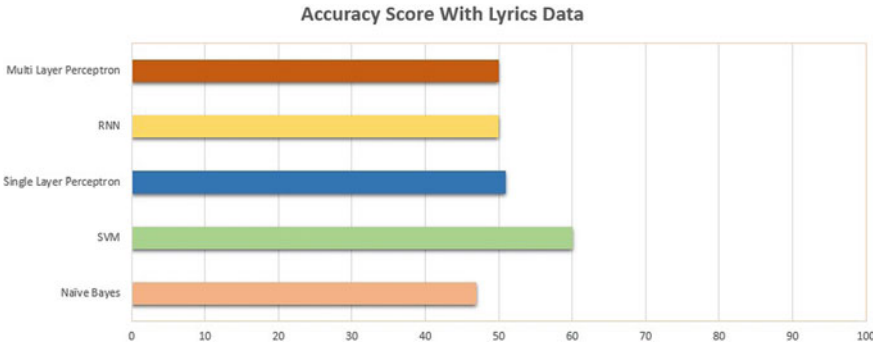


Fig. 7 Comparison of accuracy of lyrics data against different learning methodologies

Table 1 Learning measures of lyrics data through naive Bayes classifier

Parjaay	Precision	Recall	F1-score
Prakriti	0.80	0.06	0.12
Puja	0.46	1.00	0.63
Bichitro	0.00	0.00	0.00
Prem	0.50	0.06	0.10

Table 2 Learning measures of lyrics data through support vector machine

Parjaay	Precision	Recall	F1-score
Prakriti	0.59	0.61	0.60
Puja	0.63	0.88	0.74
Bichitro	0.40	0.60	0.10
Prem	0.52	0.40	0.45

Table 3 Learning measures of lyrics data through single-layer perceptron

Parjaay	Precision	Recall	F1-score
Prakriti	0.47	0.52	0.49
Puja	0.60	0.72	0.65
Bichitro	0.22	0.11	0.15
Prem	0.43	0.37	0.40

Table 4 Learning measures of lyrics data through multi-layer perceptron

Parjaay	Precision	Recall	F1-score
Prakriti	0.43	0.42	0.42
Puja	0.63	0.70	0.66
Bichitro	0.16	0.11	0.13
Prem	0.39	0.43	0.41

Table 5 Learning measures of lyrics data through RNN

Parjaay	Precision	Recall	F1-score
Prakriti	0.70	0.23	0.34
Puja	0.48	0.98	0.65
Bichitro	0.00	0.00	0.00
Prem	0.54	0.10	0.17

5.2 Audio Data

The audio features being numerous require a feature engineering to be carried out before fitting into the model. After the feature engineering through principle component analysis (PCA), the audio features were subjected for testing. We found RNN & LSTM to be two appropriate model for this purpose. The accuracy measure for audio data is quite high which is reflected in Fig. 8

The common model used for both lyrics and audio is RNN which requires a comparative study between their accuracy score. Figure 9 reflects this comparison.

Fig. 8 Accuracy of audio dataset through RNN

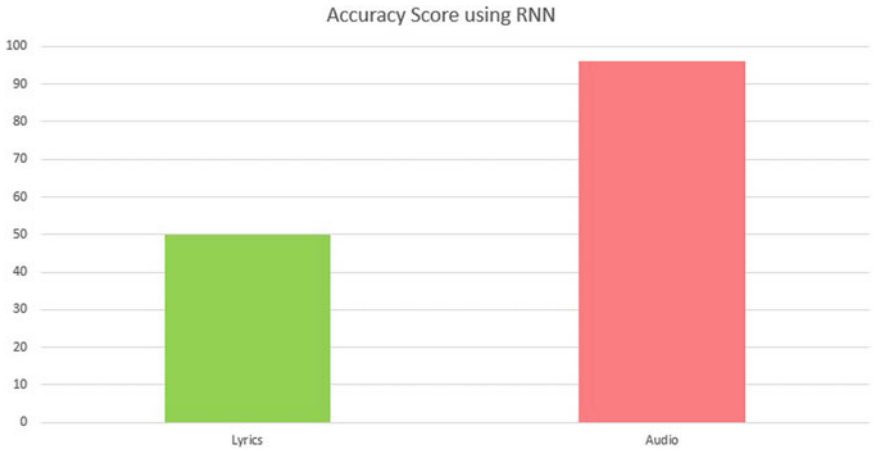


Fig. 9 Comparison of accuracy for lyrics and audio tested in RNN

6 Performance of the Classification Models

We measured the relative performance of the classification models through confusion matrices. As an example, Figs. 10 and 11 are depicted to show the performance of SVM and single-layer perceptron model, respectively, through the related confusion matrix. This present approach reveals that the machine learning and deep learning models can be used as a means for detecting the inherent semantics that is the essence of literature.

ACTUAL VALUE				
Prakriti	38	14	1	9
Puja	4	125	1	12
Bichitro	5	24	2	5
Prem	16	25	1	28
	Prakriti	Puja	Bichitro	Prem

Predicted Value

Fig. 10 Confusion matrix related to the SVM model used

ACTUAL VALUE				
Prakriti	32	21	3	6
Puja	14	102	4	18
Bichitro	4	22	4	6
Prem	18	18	7	26
	Prakriti	Puja	Bichitro	Prem

Predicted Value

Fig. 11 Confusion matrix related to the single-layer perceptron model used

7 Conclusion and Future Scope of the Work

The present work attempted to extract the dominant emotional motifs inherent to the semantics of Tagore songs. We believe that this work will create avenues to analyse the literary corpus of a language which holds the fifth position according to the Enthnologue, 2019. The analysis used here will help to find the linguistic and musical patterns to dig deep into the overall linguistic motif of Bengali language. We would like to use further deep level analysis using other advanced levels of machine learning techniques Viz. random forest, temporal CNN and attention-based models in the future. In that way, it would help to analyse and unfold the mostly implicit semantic structure within a language corpus through machine learning. Which has a wide spectrum of implementation in order to unfold the themes woven within the best literary works of a language.

References

1. Koduri GK, Indurkhya B (2010) A behavioral study of emotions in south Indian classical music and its implications in music recommendation systems. In: Proceeding of the ACM workshop on social, adaptive and personalized multimedia interaction and access, pp 55–60. ACM
2. Hu X, Downie SJ (2010a) Improving mood classification in music digital libraries by combining lyrics and audio. In: Proceeding of the 10th annual joint conference on Digital libraries. ACM, pp 159–168
3. Hu X, Downie SJ (2010b) When lyrics outperform audio for music mood classification: a feature analysis. In: Proceeding of the ISMIR, pp 619–624
4. Hu X, Downie SJ (2010) Music and mood: where theory and reality meet. In: Proceeding of the conference
5. Patra BG, Dipankar D, Bandyopadhyay S (2013) Unsupervised approach to Hindi music mood classification. In: Mining intelligence and knowledge exploration. Springer International Publishing, pp 62–69
6. Patra BG, Dipankar D, Bandyopadhyay S (2015) Music emotion recognition. In: International symposium frontiers of research speech and music (FRSM - 2015)
7. Yi-Hsuan Y, Yu-Ching L, Ya-Fan S, Homer CH (2008) A regression approach to music emotion recognition. IEEE Trans Audio, Speech Language Proc 16(2):448–457
8. Xie S, Rastogi R, Chang M (2017) Deep poetry: word-level and character-level language models for shakespearean sonnet generation

Music Genre Classification with Convolutional Neural Networks and Comparison with F, Q, and Mel Spectrogram-Based Images

Aditya Dhall, Y. V. Srinivasa Murthy, and Shashidhar G. Koolagudi

Abstract Content-based music information retrieval (CB-MIR) is the growing research field that includes many subtasks. Technology advancements could be the main reason for this. The tasks of artist identification, genre classification, instrument classification, and mood estimation are some of the important tasks of CB-MIR. In this work, an effort has been made to classify the songs into the class of genres. Ten different genres are considered with 100 clips for each genre class that are collected from the standard GTZAN dataset. Three transformations are applied in order to construct the spectrograms. They are Fourier transform-based (FS), Q transform-based (QS), and Mel frequency transform-based (MS) spectrograms. Convolutional neural networks are used for this work. AlexNet and LeNet-5 are considered for experimentation. Training accuracy is almost 100% in the case of genre classification. However, the testing accuracy is quite less when compared to the training accuracies.

Keywords AlexNet · Convolutional neural networks · Fourier-Q-MelFrequency transform-based spectrograms · LeNet-5 · Music genre classification

1 Introduction

The number of tracks is increasing tremendously due to the regular additions to the digital cloud [1]. Hence, it is essential to categorize the clips in order to access them easily. This shows the importance of specific unique identify that helps in clustering the songs into various groups. Further, it helps to locate the song easily. There are several attributes, can be called as meta-information of a music clip that includes singer, album name, title of the song, genre, mood, etc. [2, 3]. Genre is one of the important attributes that can help in effectively categorizing the clips into clusters.

A. Dhall (✉) · Y. V. Srinivasa Murthy
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu 632059, India

S. G. Koolagudi
Department of CSE, National Institute of Technology Karnataka (NITK), Surathkal, Mangalore, Karnataka 575025, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_20

Genre identification also helps in recommending songs to the listeners. It broadly helps the users to get recommended songs based on their genre interest. Hence, the task of genre identification has been taken into consideration for this work and proposed a novel approach with the support of spectrogram-based convolutioanl neural networks (CNNs) [4].

The task of genre identification is mainly helpful in two important aspects. One is to categorize and indexing the audio clips effectively. As there are millions of tracks available in digital cloud, it is highly difficult to extract the required track. In such case, genere is one attribute that helps to separate the tracks up to some extent as it can be one category. The other aspect of application could be music recommender system.

This work focuses on three types of Spectrogram images. These are F-transform, Q-transform, and Mel transform. F-transform is the most basic type of spectrogram image. Q-transform [5] and Mel transform each help in dealing with certain inefficiencies that Fourier transform suffers from. The behaviour of each of these images is different in different types of environments. While developing any type of project, the lesser the need to conduct research, the better it is for the developers as they are able to deliver their software in a shorter amount of time. This paper will study the behaviour of these three types of spectrogram images in different, well-known convolutional neural network models, which have a good track record in processing large amounts of data. Building up on this data can also open up many avenues in the field of music genre identification.

In order to conduct this study, multiple datasets need to be created. In this project, from each audio file, a spectrogram image will be extracted using the F-transform, Q-transform, and Mel transform approach. This project uses the GTZAN dataset, which is an audio dataset of one thousand songs divided equally among ten genres. These genres are *blues*, *classical*, *country*, *disco*, *hip-hop*, *jazz*, *metal*, *pop*, *reggae*, and *rock*. All these thousand songs are converted into nine different types of datasets, and three datasets per type of spectrogram are generated. The RGB datasets are then passed to the *AlexNet* convolutional neural network model in order to study their performance. Similarly, the grayscale and the normalized grayscale datasets are passed to the *LeNet-5* convolutional neural network model to compare the performance of the three types of spectrogram images. The difference in performance between grayscale and normalized grayscale images is also monitored.

The rest of the paper has been organized as follows: Sect. 2 details the literature happened for the process of music genre recognition. The technical implementation of the project has further been explained in Sect. 3. After performing all the steps in the above section, its results and observations can be found in Sect. 4 of this paper. With the help of the results, certain observations can be made that will prove to be helpful in the implementation or further research of genre identification of audio using spectrogram images. The details are given in Sect. 5.

2 Literature Review

Music genre-based classification has been introduced by the Tzanetakis and Cook in the year of 2002, after a thorough study made on the tasks of music information retrieval (MIR) [REF]. Their efforts are the base research for the significant number of publications on the task of music genre classification. The total number of publications is around 500 since 2002. However, the task of detecting the music genre is still challenging due to the many factors involved. Some notable characteristics are the uncertain definition of the genre, ambiguity in the genre, fuzzy involved while perceiving the song, etc. Moreover, the song which you feel is rock need not be a rock actually, or others may feel it belongs to the pop category. Due to these ambiguities, one can say that the complexity involved in identifying the genre of a song is quite high. Moreover, the research on music genre classification is focused on a particular dataset. There are several datasets available for the task of music genre classification. Some of the prominent datasets are GTZAN (2002) [6], Ballroom (2004) [7], ISMIR2004 (2004) [8], Latin (2008) [9], Magnatagatune (2009) [10], ISMIR2011 (2011) [11], MSD (2011) [12], MER31K (2013) [13], AcousticBrainz (2015) [14], FMA (2017) [15], Audioset (2017) [16], Kara1k (2017) [17], and MuMu (2017) [18].

The first work which was recorded is done with the two prominent classifiers called k-nearest neighbours (k-NN) and Gaussian mixture model (GMM). The accuracy obtained with these two classifiers is 60% and 61%. The features extracted from the audio signal based on acoustic analysis, and their length is equal to 30. The dataset considered for this dataset is called GTZAN, which is having 1000 clips that are collected for ten genres. Further, the dataset is popularly called GTZAN dataset. A majority of the works have been considered the same dataset for either experimentation or comparative study. However, the recent literature which was done on music information retrieval (MIR) has quoted that there are around 108 genres that are available of which, many are ambiguous.

Based on the first publication provided with the GTZAN dataset, the research of music genre classification has explored. Several features have been used with multiple classifiers and on various datasets in order to obtain better accuracies. Initial experiments are done on time-domain features such as pitch, energy, duration, timbre, and rhythm. A majority of the features are extracted from the shorter segments of length around 20 - 100 ms. Features that are obtained from the time-domain are not suitable as they are unable to discriminate the genre classes effectively. Further, frequency domain analysis is done through short-time Fourier transformation (STFT), which helps to construct the frequency spectrum. A thorough analysis is done on the spectrum in order to obtain spectral features such as spectral flux, spectral centroid, spectral roll-off, and spectral shape. Further, Mel spectrogram has been constructed, and reciprocal has been taken in order to obtain a cepstrum. Cepstral features such as Mel-frequency cepstral coefficients (MFCCs) are also used for detecting the genre. The reason for considering such features is because of their ability to mimic the human auditory system. Hence, MFCCs and their first and second-order differential

values such as Δ and $\Delta\Delta$ are used in order to improve the accuracy of genre classification. Further, analysis of the Q spectrogram, which produces Q-transform-based features, and Chroma features are also extracted to achieve better performance.

The features mentioned above are generally called as hand-crafted features as they are extracted based on signal processing techniques. Since they are extracted by analysing the signal manually, they are not reliable for getting better accuracy. Hence, the automatic feature extraction process has been introduced in order to extract the features directly from the raw signal. Sparse coding (SC) is the initial approach proposed in order to estimate the features automatically. In recent years, deep neural networks are able to identify the suitable features automatically as they are capable of mimicking the human brain. As the results obtained with sparse coding-based features are evident in recognizing the music genre, they have not been considered due to some unreliable results obtained in some research works. However, the dataset for music genre classification would be complex and large enough. Hence, some suitable and reliable classification and auto feature extraction techniques are to be used. Based on this, deep networks have been adopted in order to estimate the features automatically and to classify the category of genre effectively.

The origin of deep networks is deep belief networks (DBN), where they are initially considered in order to extract the features automatically for the various tasks like music genre and artist classification. The performance obtained with this approach is around 73.1%. When compared to the MFCCs, DBN-based features are accurate in determining the genre of a music clip. With this, a majority of the works since 2004 have adopted the DBN approach in order to estimate the features automatically. However, the reasons for their better performance when compared to MFCCs are not clear until 2013. In 2013, the internal design of DBN that helps in estimating the melody and rhythm-related features had been detailed. Melody and rhythm are suitable in classifying tasks like raga, artist, composer, mood, and genre. It could be the reason for the better performance claimed by the approach with the features extracted using DBN.

It is a known fact that the CNNs are outperforming all other traditional image processing approaches. There are several publications in the field of image processing using CNNs during the last decade. The task of music signal processing has derived the CNN approach in order to improve the accuracy of the MIR tasks as well. Since it is difficult to process the one-dimensional raw signal directly by CNN, spectrograms have been considered to feed CNNs instead of the raw signal. Spectrograms have considered as images that are to be processed by CNNs. The performance obtained with CNNs is quite good when compared to both the DBN-based features and manually extracted MFCCs as well. Hence, CNNs are used for many of the music classification tasks in order to improve accuracy. However, recurrent neural networks (RNNs) are also considered to feed the raw signal directly as the spectrograms may miss some useful information that could be useful for classifying the tasks of MIR. Some works are already using RNNs, however, not much focused on the tasks of music genre classification.

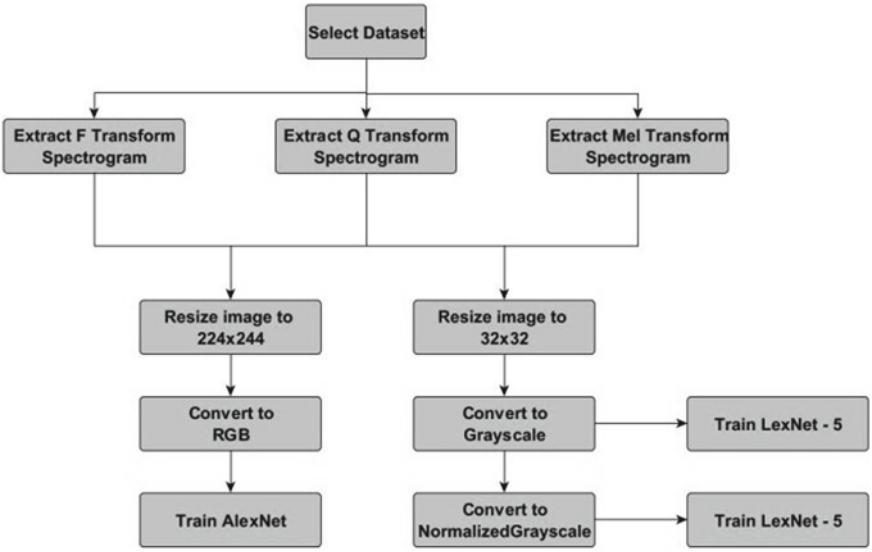


Fig. 1 Proposed flow diagram for the task of music genre classification using F, Mel, and Q spectrograms

Based on the literature mentioned above, the present authors have identified a few aspects that are useful to classify the genre accurately. One is the use of various spectrograms, and the other is the use of RNNs. In this work, an effort has been made in order to compare the performance of music genre classification using spectrograms obtained through various transformation techniques called Fourier, Q, and Mel. The spectrograms obtained using the above methods have been fed to convolutional neural networks (CNNs) that help to identify the suitable spectrogram for genre classification. The dataset considered for this work is the popular GTZAN dataset which is having 1000 songs of ten different genres namely *blues*, *classical*, *country*, *disco*, *hip-hop*, *jazz*, *metal*, *pop*, *reggae*, and *rock*. The details of the proposed methodology and the result analysis have been given in the subsequent subsections.

3 Proposed Methodology

The proposed approach has mainly two components. One is the process of constructing spectrograms, and the other is feeding them to CNN in order to compare the results. The proposed flow diagram has been clearly given in Fig. 1. A detailed explanation of each block has been given in the following subsections.

3.1 Dataset Collection

The dataset considered for this work is called GTZAN. It has a total of 1000 clips for the ten genres where 100 song clips are distributed to each genre. The clips are resampled to 44100 Hz sampling frequency and the number of bits considered for each sample is equal to 16. The total size occupied by all the clips is around 1.2 GB. Genres considered for this work are *blues*, *classical*, *country*, *disco*, *hip-hop*, *jazz*, *metal*, *pop*, *reggae*, and *rock*. A variety of clips are involved in the standard dataset of which, 65% of the data has been considered for training and the rest 35% has been considered for testing. All the clips are segmented with the frame length of 50 milliseconds prior to the generation of spectrograms.

3.2 Spectrogram Generation

Once after extracting the training and testing clips, spectrograms are constructed for every 50 milliseconds of the length. In such a way, we are constructing three different spectrograms based on Fourier and Q transformations. The three spectrograms are Fourier spectrogram (FS), Q-transform-based spectrogram (QS), and Mel transform-based spectrogram (MS). The process of constructing the spectrograms is detailed here.

3.2.1 F-Transform

There could be many single-frequency sound waves in a given frame [19, 20]. It should be analysed by converting it into a frequency domain for thorough analysis. Short-time Fourier transform, also called STFT, is one technique that helps us in capturing the individual amplitudes that an audio signal is composed of. It helps in representing the audio signal in the form of frequencies and the magnitude of each individual frequency [21].

3.2.2 Q-Transform

The constant Q-transform uses a similar concept which was used by a short-time Fourier transform (STFT) [5, 22]. In this approach, the audio signal is divided into narrow time segments, also called bins. Then, the Fourier transform is calculated for each individual bin. The difference is instead of computing a single FFT for the whole audio signal, and multiple FFTs' are calculated for the same audio signal by dividing it into different segments with constant time intervals. However, there is a glaring problem with this formula. The size of the window in the case of STFT is uniform throughout without respect to the value of the frequency. This means

that an audio signal emitted from a drum and the audio signals emitted from a flute will be treated as the same. Hence, the resolution of the signal is preserved with the Q-transform-based spectrograms when compared to STFTs. To patch this, we employ the use of constant Q-transform. This tackles the above-mentioned problems by patching two things:

1. Make sure that the bins are spaced logarithmically.
2. The size of the window is adjusted according to the central frequency. This is done to give the function and the ability to adjust the width of the bin, respectively.

3.2.3 Mel Transform

Mel transform-based spectrograms utilize the Fourier transform formula. Initially, the fast Fourier transform method is used on the audio file to compute the frequency array. Once after obtaining the frequency array, the frequencies are converted into the Mel scale. The signal is constantly changing with respect to the time, and hence, it is called a time-invariant signal. The Mel scale assumes that given a small period of time (Eg. 1–2ms), statistically speaking, the audio signal undergoes no change. According to this assumption, if the time frame is short, then there will be insufficient samples in order to get an adequate spectral estimate. In contrast, if the time frame is too long, then the signal changes too much in a single frame. With this, it is possible to discriminate between two sounds where the human ear fails in such a case. Hence, multiple bins are considered and added up in order to get an estimate of the amount of energy that exists in different frequency regions.

The STFT spectrogram extraction is performed with the help of *matplotlib* library. This library consists of functions which make plotting graphs is very easy. On the other hand, Q and Mel transform-based spectrogram images are extracted using the *LibROSA* library in Python. *LibROSA* has functions that make audio processing much easier. The sample spectrograms constructed using all three transformations have been depicted in Fig. 2.

3.3 Convolutional Neural Networks (CNNs)

This section describes the network architecture, image size, and the process of training and testing models, various models considered for this work.

3.3.1 Converting Images in Matrix Form

Initially, all the images are scaled to $224 \times 224 \times 3$. Further, the images are converted into matrix forms in order to train their respective models. We have considered AlexNet as it supports the images of three channels (RGB). However, LeNet - 5 uses

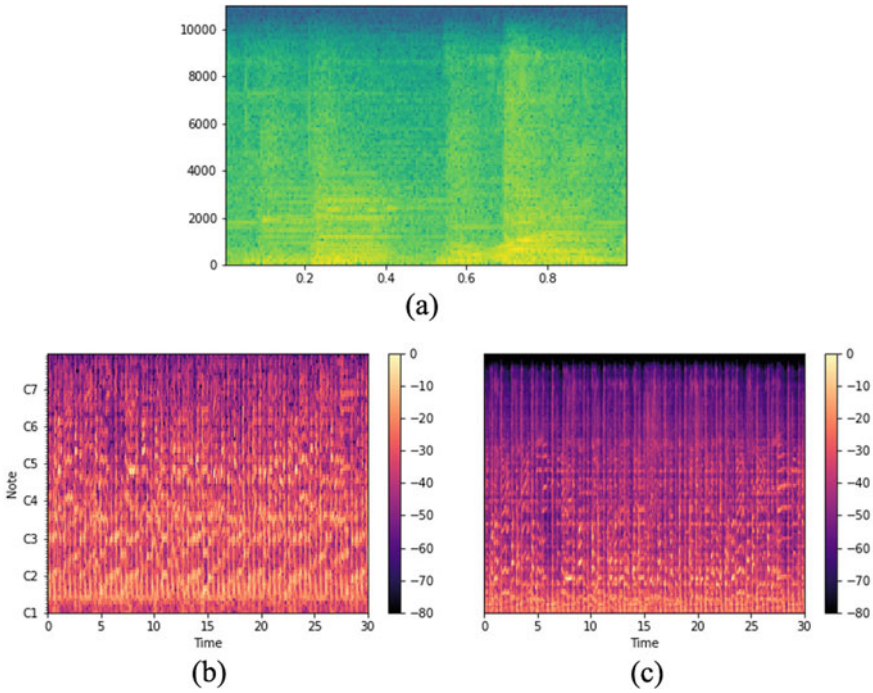


Fig. 2 Sample spectrograms obtained using STFT, Q, and Mel transformations. **a** F-transform based, **b** Q-transform based, and **c** Mel-transform based spectrograms

a single channel while training. This means that each image from the three datasets will be converted into a grayscale matrix of dimensions $32 \times 32 \times 1$. Now, each type of spectrogram has been converted into a grayscale matrix, and it can now be normalized. This is achieved by dividing the entire matrix by 255. Based on this, one can say that each type of spectrogram has two datasets. So, in total, there are nine (3 from AlexNet + 6 from LeNet-5) datasets generated for this study.

3.3.2 Training Models

AlexNet was first introduced in September 2012 in the ImageNet competition, where it managed to achieve a top-5 error of 15.3%, which had beat the next best model by 10.8%. It can be considered as a notable achievement. This was credited to the depth of the model. However, it would be computationally expensive to implement. It needs high computational power in order to handle the complexity, preferably graphics processing units (GPUs). AlexNet model can be divided into five convolutional layers and three dense layers [23].

LeNet-5 is one of the earliest CNN models, introduced in the year 1998 and proposed by Yann LeCun. However, as there was a lack of efficient hardware in 1998,

it never gained a lot of recognition until very recently. It is a small and straightforward model and can even run on a simple modern-day CPU.

Keras in Python has been used to implement each of the models. Keras is a sub-library of the Tensor-flow library in Python. Before training the models, each dataset is divided into training and testing data. Out of 1000 images in each dataset, 350 images are randomly chosen and will be used to test the validation accuracy of the model. Their respective datasets are now passed into each of these models in order to study the performance of each type of spectrogram in different environments. After training has been completed, the accuracy, validation accuracy, accuracy graph, and loss graph of each spectrogram are used to compare them.

4 Result Analysis

This section details the experimental results obtained using the proposed approach, and the analysis made out of it.

Prior to the experimentation, the dataset has been divided into 65% and 35% as training and testing sets, respectively. The experimentations are done in phase wise using both Alex Net and LeNet-5. We have utilized the variations of spectrograms like three-channel RGB and single-channel grey coloured spectrograms as well. Moreover, spectrograms obtained using three transformation techniques are considered for experimentation. The results obtained with the specified information have been depicted in Table. 1. The table is having four columns where the first column represents the spectrogram type considered having values like Fourier spectrogram (FS), Q-transform-based spectrogram (QS), and Mel spectrogram (MS). The second column indicates the results obtained using AlexNet for which RGB spectrograms have been fed. The third and fourth columns represent the results obtained using LeNet-5.

Table 1 Training and test accuracies obtained using AlexNet (RGB), LeNet-5 (grey normalized and grey). Three spectrograms such as FS, QS, and MS are used. Accuracy → Training Accuracy

Spectrogram Type	AlexNet (RGB)	LeNet 5	
		Grayscale Normalized	Grayscale
F	Accuracy = 62.46% Test Accuracy = 10.00%	Accuracy = 44.46% Test Accuracy = 41.43%	Accuracy = 48.92% Test Accuracy = 34.57%
Q	Accuracy = 91.38% Test Accuracy = 22.29%	Accuracy = 71.23% Test Accuracy = 52%	Accuracy = 100% Test Accuracy = 41.71%
Mel	Accuracy = 80.00% Test Accuracy = 50.86%	Accuracy = 68.31% Test Accuracy = 52%	Accuracy = 100% Test Accuracy = 39.71%

Further, the greyscale images are categorized into two sets, one with normalized greyscale images, and the other is raw ones.

This paragraph is discussing the results obtained using AlexNet on all the three classes of spectrograms. It is found that the accuracy obtained with F-transform-based spectrograms is very less in all the cases when compared to other spectrograms. It is around 62.46% which is obtained with AlexNet on RGB spectrograms. The test accuracy obtained with F-transform-based spectrograms is around 10%. In the case of AlexNet, Q-transform-based spectrograms are giving better accuracies with an accuracy of 91.38%. However, the test accuracy is quite less and around 22.29%. Mel spectrogram-based AlexNet is giving an accuracy of 80%. The test accuracy is around 50.86%. Hence, Mel spectrograms are found to be good with the GTZAN dataset when compared to the other two spectrograms.

In the case of LeNet-5, the performance obtained with the grey spectrograms is found to be accurate when compared to the normalized grey spectrograms. Initially, normalized greyscale spectrograms are used, and the accuracies are 44.46%, 71.23%, and 68.31% for FS, QS, and MS, respectively. The test accuracies for the same are 41.30%, 52%, and 52%. One can say that the Q-transform-based spectrograms are

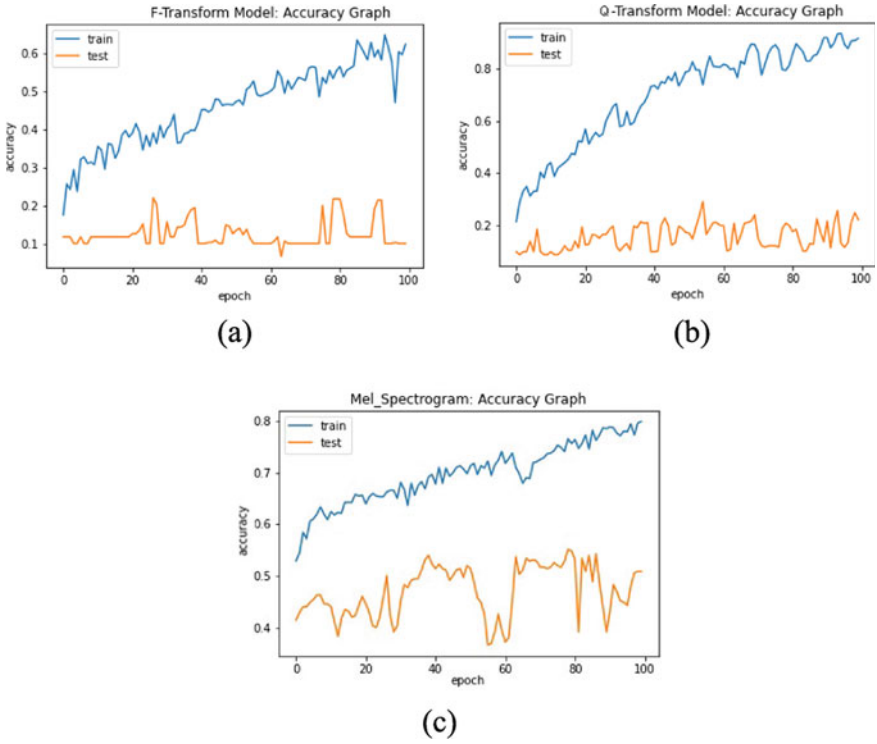


Fig. 3 AlexNet (RGB): Train and test accuracy graphs obtained for F-transform-based, Q-transform-based, and Mel transform spectrograms. **a** FS, **b** QS, and **c** MS

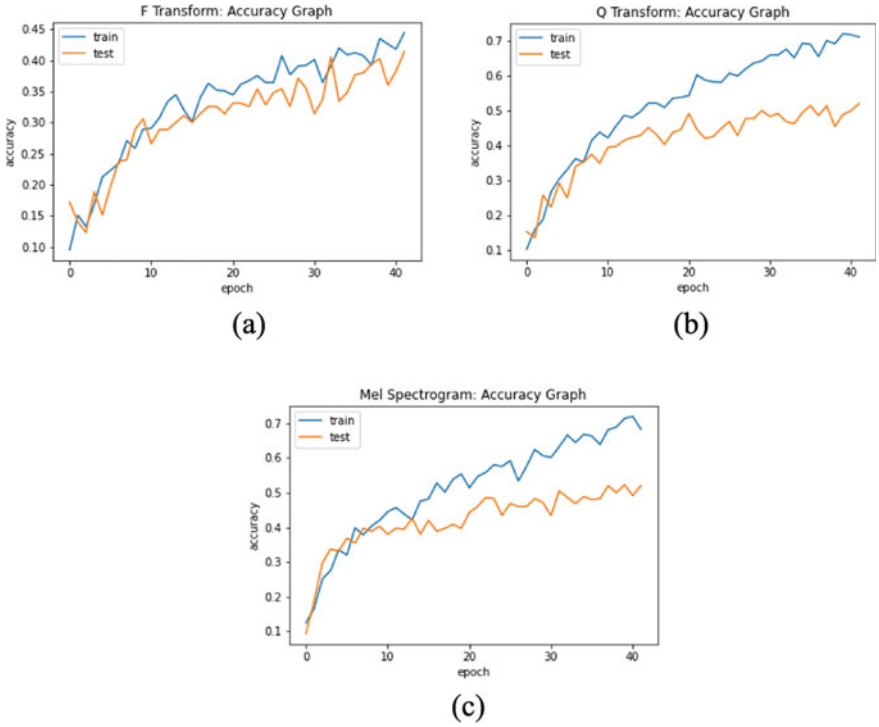


Fig. 4 LeNet-5 (Grey Normalized): Train and test accuracy graphs obtained for F-transform-based, Q-transform-based, and Mel transform spectrograms. **a** FS, **b** QS, and **c** MS

performing better in this case. Both the train and test accuracies are accurate in this case. Further, raw greyscale spectrograms are used for experimentation. The train accuracies are 48.92%, 100%, and 100%, and 34.57%, 41.71%, and 39.71% are the test accuracies for FS, QS, and MS, respectively, of which, Q transform-based spectrograms are giving better performance in both the train and test accuracies. Though the accuracy is 100% with Q and Mel spectrograms, the test accuracy is very less which can result in underfitting issues.

The graphs obtained with the AlexNet (RGB spectrograms), LeNet-5 (grey and normalized grey spectrograms) are shown in Fig. 3, 4, and 5. Figure 3 is obtained using AlexNet, and three columns represent FS, QS, and MS. The rest 4 and 5 are for grey normalized and grey spectrograms. Detailed training and testing accuracy graphs are given here.

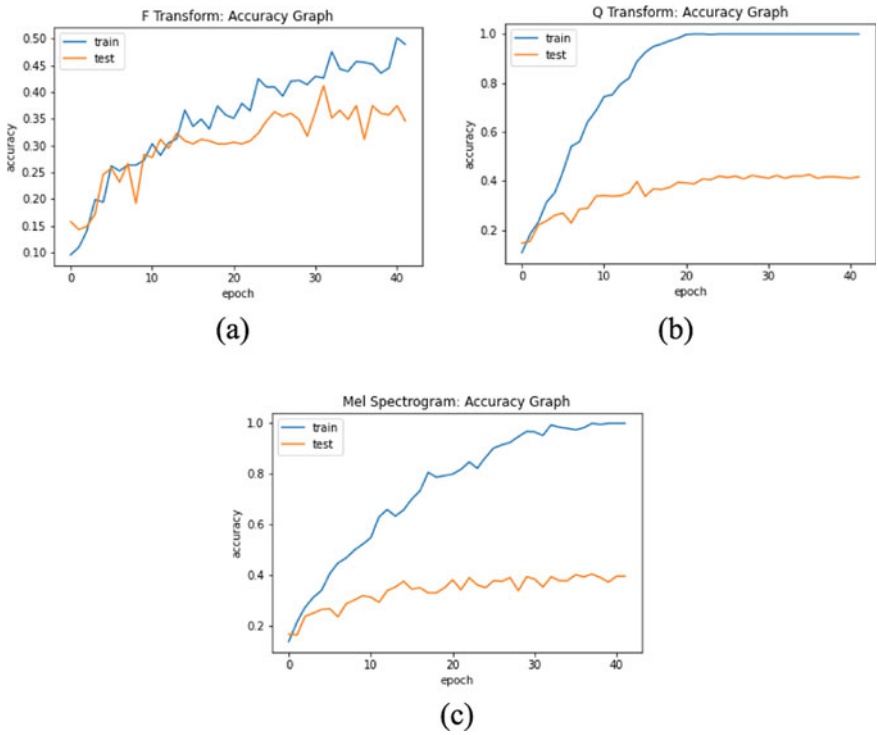


Fig. 5 LeNet-5 (Grey): Train and test accuracy graphs obtained for F-transform-based, Q-transform-based, and Mel transform spectrograms. **a** FS, **b** QS, and **c** MS

5 Conclusion and Future Work

The task of music genre classification using spectrogram-based CNNs has been considered for this work. Since the hand-crafted features are not giving an expected accuracy, raw spectrograms have been considered for the genre classification. Moreover, Fourier transform-based spectrograms are becoming outdated and not carrying much useful information from the complex music signals. Hence, the spectrograms based on Q and Mel transformations have been constructed as they carry out the high resolution with respect to timbral information. As expected, the results obtained with the Q transformation and Mel spectrograms are accurate when compared to the results obtained using Fourier transform-based spectrograms. The accuracies are appropriate for the considered GTZAN dataset. Average accuracy of 100% in case of training and around 56% with the testing has been obtained.

Since the popular CNNs are outperforming when compared to the hand-crafted features, it is essential to do certain experimentation on the real dataset. Moreover, RNNs are also more suitable as they are capable of extracting the features directly from the raw signal. In this case, we need not put any effort even to construct the spectrograms. Hence, our future work will throw a light on the aspects to compare the accuracies of various available datasets using CNNs and RNNs.

References

1. Srinivasa Murthy YV, Koolagudi SG (2018) Content-based music information retrieval (cb-mir) and its applications toward the music industry: a review. ACM Comput Surv (CSUR) 51(3):1–46
2. Srinivasa Murthy YV, Koolagudi SG (2015) Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations. In: 2015 IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, pp 1271–1276
3. Srinivasa Murthy YV, Koolagudi SG (2018) Classification of vocal and non-vocal segments in audio clips using genetic algorithm based feature selection (gafs). Expert Syst Appli 106:77–91
4. Murthy YVS, Jeshventh TKR, Zoeb M, Saumyadip M, Shashidhar GK (2018) Singer identification from smaller snippets of audio clips using acoustic features and dnns. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, pp 1–6
5. Thomas M, Srinivasa Murthy YV, Koolagudi SG (2016) Detection of largest possible repeated patterns in indian audio songs using spectral features. In: 2016 IEEE Canadian conference on electrical and computer engineering (CCECE). IEEE, pp 1–5
6. Sturm BL (2013) The gtzan dataset: its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint* arXiv:1306.1461
7. Marchand U, Peeters G (2016) The extended ballroom dataset
8. Cano P, Gutiérrez EG, Gouyon F, Boyer H, Koppenberger M, Ong BS, Serra X, Streich S, Wack N et al (2006) Ismir 2004 audio description contest
9. Hu Y, Ogihara M (2012) Genre classification for million song dataset using confidence-based classifiers combination. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, pp 1083–1084
10. Wolff D, Weyde T (2012) Adapting similarity on the magnatagatune database: effects of model and feature choices. In: Proceedings of the 21st international conference on world wide web, pp 931–936
11. Marques CM, Guilherme IR, Nakamura RYM, Papa JP (2011) New trends in musical genre classification using optimum-path forest. In: ISMIR, pp 699–704
12. Schreiber H (2015) Improving genre annotations for the million song dataset. In: ISMIR, pp 241–247
13. Chen P-I, Liu J-Y, Yang Y-H (2015) Personal factors in music preference and similarity: user study on the role of personality traits. In: Proceedings of international symposium on Computer Music Multidisciplinary Research (CMMR)
14. Bogdanov D, Porter A, Schreiber H, Urbano J, Oramas S (2019) The acousticbrainz genre dataset: multi-source, multi-level, multi-label, and large-scale. In: Proceedings of the 20th conference of the International Society for Music Information Retrieval (ISMIR 2019): 2019 Nov 4-8; Delft, The Netherlands.[Canada]: ISMIR. International Society for Music Information Retrieval (ISMIR)
15. Defferrard M, Benzi K, Vandergheynst P, Bresson X (2016) Fma: a dataset for music analysis. *arXiv preprint* arXiv:1612.01840
16. Bahuleyan H (2018) Music genre classification using machine learning techniques. *arXiv preprint* arXiv:1804.01149

17. Bayle Y, Maršík L, Rusek M, Robine M, Hanna P, Slaninová K, Martinovic J, Pokornỳ J (2017) Kara1k: a karaoke dataset for cover song identification and singing voice analysis. In: 2017 IEEE International Symposium on Multimedia (ISM). IEEE, pp 177–184
18. Oramas S, Nieto O, Barbieri F, Serra X (2017) Multi-label music genre classification from audio, text, and images using deep features. *arXiv preprint* arXiv:1707.04916
19. Srinivasa Murthy YV, Koolagudi SG, Swaroop VG (2017) Vocal and non-vocal segmentation based on the analysis of formant structure. In: 2017 ninth International Conference on Advances in Pattern Recognition (ICAPR). IEEE, pp 1–6
20. Sharma R, Srinivasa Murthy YV, Koolagudi SG (2016) Audio songs classification based on music patterns. In: Proceedings of the second international conference on computer and communication technologies. Springer, Berlin, pp 157–166
21. Oppenheim AV (1970) Speech spectrograms using the fast fourier transform. IEEE Spectrum 7(8):57–62
22. Thomas M, Jothish M, Thomas N, Koolagudi SG, Srinivasa Murthy YV (2016) Detection of similarity in music files using signal level analysis. In: 2016 IEEE Region 10 Conference (TENCON). IEEE, pp 1650–1654
23. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556

A Simple Statistical Comparison of Bageshree and Bhimpalasi

Pranjala Shukla and Soubhik Chakraborty

Abstract From the genesis of Indian classical music till the present day, there have been various modifications and transitions from its perspective. Indian classical music, which originated from the historically illustrious Indian culture, holds its importance worldwide from primeval times. This paper performs a simple statistical comparison of two ragas of the Kafi thaat—Bageshree and Bhimpalasi. Raga Bageshree, on one hand, is well known for its ability to relax the senses and calm the mind. On the other hand, Bhimpalasi brings together a joyful and tender mood. Although these two ragas originate from the same thaat, the statistical analysis shows some interesting results. These results when calculated for multiple ragas, in the future, can be used as an input to extract the raga information from any composition.

Keywords Probability · Coefficient of variation · Raga · Melody length · Thaat

1 Introduction

Classical music to date is preeminent among all other genres because it holds the core elements of the soul and raw emotions unaffected by human interference within it. Classical composers wrote enduring masterpieces that were meant to last forever and will always be pertinent to the human race.

Classical music has widely demonstrated its intelligence traits. This quality of classical music has been part of many studies. Researchers have shown that classical music has lasting effects on the brain. It can also help in enhancing psychological health. Classical music furnishes endless benefits which continue to unfold with the multiple studies. This genre is made up of melodic compositions (called *bandish*) which were written ages ago. Now, they continue to serve as a historical collection.

In this paper, 'Kafi thaat' is taken into account for statistical analysis because it is one of the ancient thaats and its scale is considered to be a natural scale. Two ragas

P. Shukla
Department of Computer Science and Engineering, BIT Mesra, Ranchi 835215, India

S. Chakraborty (✉)
Department of Mathematics, BIT Mesra, Ranchi 835215, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_21

of Kafi thaat, namely Bageshree (or Bageshri) and Bhimpalasi or Bhimpalashree), are considered for this study. A raga is a melodic framework that defines the usage of each note and also describes the dominance of a pitch over others. This alteration results in distinct ragas with their different moods. A musical scale called thaat is a classifying factor of ragas. For further knowledge and compositions of ragas, we refer to classical music books by Bhatkhande [1–3].

For applications of statistics in western music analysis, readers can refer to the book by Beran [4] and for Indian music analysis, the book by Chakraborty et al. [5] can be consulted. For musical signal processing with special emphasis on Hindustani ragas, readers can refer to Datta et al. [6].

The concepts and formulae used in this study are elaborated in the following section, namely methodology. These formulae are used in the calculations of experimental results. The probability of notes is calculated for both the ragas. Further melody lengths in each raga composition are used to calculate the mean, standard deviation, and coefficient of variation. Conditional probability is calculated for the notes of both the ragas to depict the dependence of one note on another. Further, based on these results, multiple discussion points are identified, which helps in moving forward with the conclusions of this study.

2 Methodology

For each raga, the data for all the 12 notes in an octave (Sa to Shudh Ni) is gathered. The probability and information content of each note of both the ragas are calculated. Further mean, standard deviation (SD), and coefficient of variation (CV) are determined for the melody lengths of each raga. Melody length can be understood as the number of notes in a line or melody. Melody is a linear series of notes that can be perceived as a single entity (e.g., one line of a song or, in our case, each line of a composition).

Probability of a note. It is defined as the number of favorable outcomes divided by the total number of outcomes. In our case, it is the ratio of the number of occurrences of a note to the total number of notes in a raga composition. The occurrence of notes is taken into account for this calculation and not the length of notes.

Information content. It evaluates the informativeness of a note. The information content of the occurrence of a particular note (event E) is defined as

$$I(E) = -\log_2(P(E)) \tag{1}$$

It is inversely proportional to the probability of a note. This means for a note with lesser probability, the surprise factor would be more on its appearance. To sum up, though the notes with greater probabilities are important, the ones with lesser probability cannot be discarded either, because of the element of surprise that appears

with their occurrence. This probability and surprise element of notes differ from raga to raga.

We have calculated the mean, standard deviation (SD), and coefficient of variation (CV) of the melody lengths of both raga compositions. Mean measures average, SD measures dispersion, and CV measures consistency in numerical data.

Mean. Mean here would be used as a measure of the average of melody lengths. The mean of m observations is defined as

$$\text{Mean} = \left(\sum y\right)/m \tag{2}$$

Standard Deviation. It is the measure of the variation of melody lengths (in our case), calculated as the positive square root of variance. Standard deviation (SD) of m observations is defined as

$$\text{SD} = \sqrt[+]{\frac{\sum(y - \text{mean})^2}{m}} = \sqrt[+]{\frac{\sum y^2}{m} - \left(\frac{\sum y}{m}\right)^2} \tag{3}$$

Coefficient of Variation. It is the measure of the variation of melody lengths around mean, calculated as

$$\text{CV} = \text{SD}/\text{Mean} \tag{4}$$

The lesser the CV, the greater is the consistency in the data.

Chi-Square Goodness-of-fit test: It is a statistical hypothesis test for goodness of fit of a probability model that is fitted to the numerical data. Here, we shall use it to test whether the melody lengths are uniform; y distributed over the lines of the raga composition. Chi-square denoted by χ^2 (see the remark below its definition) here would be computed as:

$$\chi^2 = \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} \tag{5}$$

Which follows chi-square distribution with $(k - 1)$ degrees of freedom, k being the number of classes. Here, k would be the number of lines in the composition. Obs is the actual or observed melody length of a line in a composition. Exp is the expected melody length or mean of all the melody lengths in the raga composition, which will be fixed for every line on presuming uniform model [7]. We shall test whether the presumption of uniform model is true or false.

Remark Chi-square (definition): Square of a standard normal variate is a chi-square variate with one degree of freedom. The mean and variance of a chi-square distribution with n degrees of freedom are n and $2n$, respectively. The distribution is positively skewed and possesses additivity property, i.e., the sum of independent

chi-square variates is again a chi-square variate with degrees of freedom as sum of the degrees of freedom of the individual variates appearing in the sum.

3 Experimental Results

Table 1 shows the frequency of occurrence (probability) of each note in both the ragas of Kafithaat. Table 2 conveys the surprise element, i.e., information content of each note in both raga compositions under consideration. Letter K beside the note stands for flat note also known as Komal note. Table 3 displays the length of melodies in both raga compositions which is used for further calculations in Tables 4 and 5. While in Table 4, the mean and SD of the melody lengths for both the ragas are calculated, these values are further used to calculate the coefficient of variation of melody lengths. Calculated chi-square values for Bageshree and Bhimpalasi ragas are displayed in Table 5. Figure 1 illustrates the graph of standard deviation versus mean of melody lengths for the two ragas. The bar diagram in Fig. 2 compares the coefficient of variation of melody lengths of the ragas Bageshree (also spelt as Bageshri by some authors) and Bhimpalasi (also spelt as Bhimpalashree by some authors).

As mentioned earlier, the chi-square goodness-of-fit test, tabulated in Table 5, has $(k - 1)$ degrees of freedom. One degree of freedom is lost because of the linear

Table 1 Probability of notes of Bageshree and Bhimpalasi

	Bhimpalasi	Bageshree
Sa	32/112	24/93
Re	06/112	10/93
Ga(K)	15/112	12/93
Ma	25/112	10/93
Pa	14/112	03/93
Dha	03/112	18/93
Ni(K)	17/112	16/93

Table 2 Information content of Bageshree and Bhimpalasi

	Bhimpalasi	Bageshree
Sa	1.81	1.95
Re	4.22	3.22
Ga(K)	2.89	2.95
Ma	2.16	3.22
Pa	3	4.96
Dha	5.23	2.37
Ni(K)	2.72	2.57

Table 3 Table of melody lengths of Bageshree and Bhimpalasi ragas

	Bhimpalasi	Bageshree
LINE 1	25	18
LINE 2	25	17
LINE 3	21	19
LINE 4	20	20
LINE 5	21	19

Table 4 Table of mean, SD, and CV of melody lengths of Bageshree and Bhimpalasi

	Bhimpalasi	Bageshree
Mean	22.4	18.6
SD	2.15	1.01
CV	0.09	0.05

Table 5 Summary of the chi-squared test of Bageshree and Bhimpalasi (details omitted)

	Bhimpalasi	Bageshree
Chi-square value	0.010	0.003

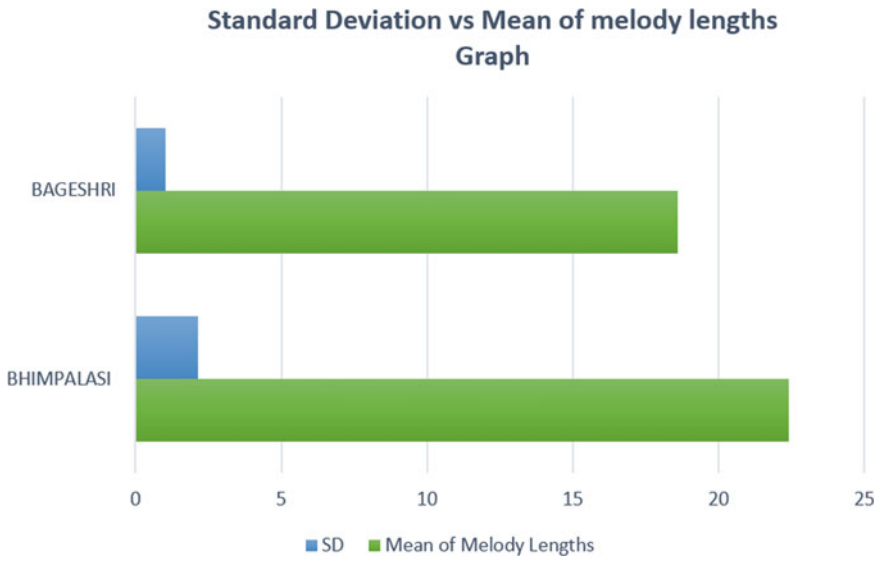


Fig. 1 A plot of std. deviation versus mean of melody lengths

restriction that the sum of observed class frequencies must be equal to that of expected class frequencies. For this study, k is considered to be the number of melodies (lines) in the considered raga composition of Bageshree and Bhimpalasi ragas. From the

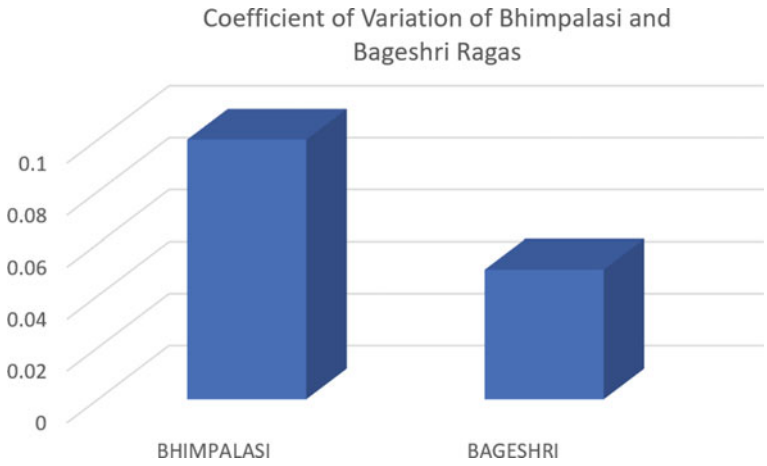


Fig. 2 A plot of coefficient of variation of different ragas of Kalyan Thaat

data in Table 3, $k = 5$. If the calculated chi-square value is less than the calculated chi-square value at 4, i.e., $(k - 1)$ degrees of freedom and 5% level of significance, we may accept the null hypothesis H_0 at 5% level otherwise the null hypothesis would be rejected. In the musical composition of both the ragas, if the length of melodies is evenly (uniformly) distributed, it will be deemed as our null hypothesis. The consideration for alternative hypothesis H_1 is that the distribution of melody lengths is uneven (non-uniform).

4 Discussion

The unconditional probability calculated in Table 1 elaborates on the importance of each note in both the ragas.

For Raga Bhimpalasi, the order of notes with lowest to highest importance is: dha < re < pa < ga(k) < ni(k) < ma < sa.

For Raga Bageshree, the order of notes with lowest to highest importance is: pa < ma = re < ga(k) < ni(k) < dha < sa.

For both the Ragas, Vadior most important note is Ma and Samvadi note is Sa. From the above order, it can be stated that though Vadinote plays an important role in ragas expression [8] but it may not be the one which has the highest number of occurrences (in both the raga compositions, musical note Sa has the highest probability but as per music theory on Hindustani ragas, it is the second most important note).

A moment's reflection on Table 2 suggests that the highest surprise element comes with the appearance of note Dhain Bhimpalasi and note Pa in Bageshree. It is musically known that these are the weak notes in the corresponding ragas. Hence, when

they arrive, they bring amazement. Also, a comparison of note phrases shows Ma-Dha-Ni(K)-Sa to be a prominent note combination in Bageshree as opposed to the phrase Ma-Pa-Ni(K)-Sa in Bhimpalasi. This also marks a musical difference between the two ragas. Another typical note phrase Ma-Pa-Dha-Ga(K) in Bageshree was absent in the Bhimpalasi composition which is quite expected musically [9, 10].

Smaller values of SD in Table 4 can be used to conclude that the melody lengths are not distributed widely. While mean melody length is greater for Bhimpalasi as compared to that of Bageshree, the SD of the former is more than twice that of the latter implying greater variation. We know lesser the CV, the greater is the preciseness (consistency). The melody lengths of Bageshree's composition are consequently more consistent as compared to Bhimpalasi's one. In fact, the % CV obtained by multiplying the CV by 100 is 9% for Bhimpalasi, nearly double against 5% for Bageshree.

Tabulated chi-square at a 5% level is 9.488 at 4 degrees of freedom. Both the calculated chi-square values in Table 5 are less than the tabulated value of 9.488. Hence, the calculated values support the null hypothesis that the melody lengths in both the Raga compositions are evenly distributed.

From the transition probability matrices in Tables 6 and 7, it can be seen, for

Table 6 Conditional probabilities of Bhimpalasi

	Sa	Re	Ga (K)	Ma	Pa	Dha	Ni (K)
Sa	12/31	0/31	3/31	3/31	1/31	0/31	12/31
Re	5/6	1/6	0/6	0/6	0/6	0/6	0/6
Ga (K)	1/15	5/15	0/15	9/15	0/15	0/15	0/15
Ma	0/25	0/25	12/25	7/25	6/25	0/25	0/25
Pa	2/14	0/14	0/14	6/14	2/14	0/14	4/14
Dha	0/3	0/3	0/3	0/3	3/3	0/3	0/3
Ni (K)	11/17	0/17	0/17	0/17	2/17	3/17	1/17

The row head gives the present note and the column head gives the next note

Table 7 Conditional probabilities of Bageshree

	Sa	Re	Ga (K)	Ma	Pa	Dha	Ni (K)
Sa	10/23	3/23	0/23	2/23	0/23	2/23	6/23
Re	7/10	3/10	0/10	0/10	0/10	0/10	0/10
Ga (K)	0/12	4/12	6/12	2/12	0/12	0/12	0/12
Ma	0/10	0/10	6/10	0/10	2/10	1/10	1/10
Pa	0/3	0/3	0/3	0/3	0/3	3/3	0/3
Dha	3/18	0/18	0/18	6/18	1/18	3/18	5/18
Ni (K)	4/16	0/16	0/16	0/16	0/16	8/16	4/16

The row head gives the present note and the column head gives the next note

example, that probability of the next note to be Dhain Bageshree, given that the current node is Ma is 1/10 but if the current node is Ni (K), the probability of Dha becomes 8/16. The conditional probability of a note thus depends on the note previously occurred. The unconditional probability of a note in a raga is not equal to the conditional probability of that particular note, given the previous note. That is, notes are dependent. Observe that Sa being the last note in both the raga sequences, we have to subtract one from the total occurrences of Sain the note sequence as we have no information of the next transition for the last note while computing the conditional probabilities. This also makes the row sum to be unity which it should be.

5 Conclusion

We have performed a simple statistical comparison-based study between Bageshree and Bhimpalasi ragas which uses the same set of notes. The intention behind selecting both ragas of the same thaat lies in demonstrating the variation in the probability of a note from raga to raga. The analysis clearly shows the relative importance of the permissible notes in the ragas. The conditional probability (of the next note, given the current one) is different from the unconditional probability of the current note. This proves that music is not random, besides it incorporates multiple types of patterns, be it patterns in time, in pitch, etc. Other features like note duration, interonset interval, and pitch movement between the notes can throw light on stay notes, presence of rhythm, and transitory information, respectively, from actual recordings although the data will vary from artist to artist. This study can be elaborated for multiple ragas of different thaat; the observations can be used to depict the raga of any composition.

References

1. Bhatkhande PVN (2005) Hindustani SangeetPaddhatiKramikPustak Malika Part-III. SangeetKaryalaya, Hathras (in Hindi)
2. Bhatkhande PVN (2006) Hindustani Sangeet Paddhati Kramik Pustak Malika Part-IV. Sangeet Karyalaya, Hathras (in Hindi)
3. Bhatkhande PVN (2002) Hindustani Sangeet Paddhati Kramik Pustak Malika Part-II. Sangeet Karyalaya, Hathras (in Hindi)
4. Beran J (2004) Statistics in musicology. Chapman & Hall, New York
5. Chakraborty S, Mazzola G, Tewari S, Patra M (2014) Computational musicology in Hindustani music. Springer International Publishing, Switzerland
6. Datta AK, Solanki SS, Sengupta R, Chakraborty S, Mahto K, Patranabis A (2017) Signal analysis of Hindustani classical music. Springer Nature Singapore
7. Rohtagi VK, Saleh AKME (2015) An introduction to probability and statistics. Wiley

8. Vasant Sangeet (2017) Visharad. Hathras, Sangeet Karyalaya
9. Ringe AVR (2020) Tanarang, (Online). Available: http://www.tanarang.com/english/bageshree_eng.htm. Accessed Sept 2020
10. Ringe AVR (2020) Tanarang, (Online). Available: http://www.tanarang.com/english/bheempalasi_eng.htm. Accessed Sept 2020

Sound Analysis and Cognition

Perception of Speech and Music, Cognition and Brain Studies, Hearing and Cognition, Emotion

Effects of Musical Training on Auditory Spatial Processing Abilities: A Psychoacoustical and Perceptual Study

Bhoomika and Kavassery Venkateswaran Nisha

Abstract The present study aimed to determine the effect of musical training on binaural and spatial processing abilities in participants aged 18–30y. A test battery consisting of the binaural processing (interaural time: ITD and interaural level difference:ILD thresholds) and spatial processing (virtual acoustic space identification test: VASI & speech spatial and qualities in Kannada: SSQ-K questionnaire) measures were carried out on two groups of listeners (Group I: Musicians $N = 15$, Group II: Non-Musicians $N = 15$). The stimuli used in all the psychoacoustic tests were 250 ms white noise bursts, while for SSQ-K participants gave subjective ratings for 14 questions. The stimuli were routed via headphones using MATLAB (binaural processing) and paradigm interfaces (VASI). Statistical analyses using MANOVA and Mann–Whitney U showed musicians performed significantly better ($p < 0.05$) than their age-matched non-musician counterparts in all the tests. The findings of the study highlight the positive outcomes of musical training in fine-refinement of binaural and spatial processing skills.

Keywords Binaural processing · Spatial processing · Musicians

1 Introduction

The relationship between musical training and auditory perception abilities is complicated, which manifests itself as the response of musical training. The role of musical training has been extensively studied in terms of many psychoacoustic skills such as pitch, time [1], and loudness discrimination scores [2], with musicians showing superior skills on all measures than age-matched non-musicians. The enhanced skill of musicians in detecting subtle changes in pitch, time, and intensity aids in auditory stream segregation [3, 4]. Better performance of musicians in auditory stream segregation task reflects their improved ability to process binaural and spectral cues. The use of binaural and spectral cues is indicated in auditory spatial processing, which

Bhoomika · K. V. Nisha (✉)
Department of Audiology, All India Institute of Speech and Hearing, Naimisham Campus, Manasagangothri, Mysore 570006, India
e-mail: nishakv@aiishmysore.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_22

is the ability to detect and monitor the sounds located in 3D space. Spatial acuity is governed by three core aspects: temporal correlates [interaural time difference—ITD] intensity correlates [interaural level difference—ILD] and spectral cues [virtual auditory space identification—VASI] [5] .

Apart from salient advantages of musical training on for fine perception of pitch [6] and level differences [7], studies have established superiority of musicians on temporal tasks too. Donai et al. [8] administered gaps-in-noise (GIN) test in 14 musicians and 14 non-musicians. The musicians in his study exhibited lower gap detection threshold (GDT) threshold (3.5 ± 0.76 ms) compared to the non-musician group 5.7 ± 1.14 ms), suggestive of musical training benefits. On similar lines, Rammsayer and Altenmüller [9] found superior performance of musicians on temporal discrimination tasks, rhythm perception, and auditory fusion. Musical training benefits were transferable to other tasks, such as visual and cognitive processing.

The positive outcomes of musical training on temporal processing were found to be extendable to auditory and visual memory, and cognition too [10–13]. Hansen et al. [11] administered the memory scale test [14], digit span test, and the Musical Ear Test (MET) [15] on a group of 20 non-musicians, 20 amateur musicians, and 20 expert musicians. The findings of the study revealed that expert musicians significantly outperformed non-musicians on the digit span test, suggestive of better working memory in the musician group. These memory-related effects of musical training were generalizable to auditory cognition, too [16]. Helmbold et al. [12] compared 70 musicians and 70 non-musicians for primary mental abilities (verbal comprehension, word fluency, space, flexibility of closure, perceptual speed, reasoning, number, and memory) and found that musicians have better memory and reasoning skills. On similar lines, Schellenberg [16] reported a positive correlation between music skill, working memory and IQ.

Although numerous literature reports discuss the superiority of musicians on temporal, memory, and cognitive skills, the effects of musical training on spatial processing are relatively scantly researched. Out of the few reports available in the literature, Kaplanis and Van Velzen [17] reported that musicians demonstrated significantly higher performance on space perception in reverberant environments, especially in the vertical distance discrimination. The higher spatial acuity in the vertical plane is indicative of enhanced ability to process intricate spectral cues [5] in musicians. However, this attempt faces a shortcoming on integrating all three core aspects (temporal, intensity, and spectral), which are very important aspects of spatial processing. While ITD and ILD account for time and level variations, they largely ignore the role of spectral cues in spatial processing. Inclusion of spectral cues along with ITD and ILD in the VASI test can be used to track musical training-related changes in spatial processing. Further, a comprehensive evaluation of spatial processing abilities in musicians and comparison of same with the control group (no formal musical training) will help in identifying the possible mechanisms (temporal processing evaluated using ITD thresholds, intensity processing evaluated using ILD thresholds, and spectral processing assessed on VASI test) for auditory spatial advantage seen in musicians.

1.1 Need for the Study

The current investigation addresses the need to understand musical training-related differences in spatial acuity in normal-hearing listeners through the inclusion of battery of psychoacoustic tests (ITD, ILD, and VASI) and subjective assessment (spatial sub-section of SSQ-K). In the present study, we aim to explore and quantify the benefits of musical training (if any) by comparing the performance of musicians and non-musicians on auditory spatial processing tests.

1.2 Aim and Objectives

The present aimed to quantify the influence of musical training on auditory spatial acuity reliant on different cues in isolation (ITD, ILD), in combination (ITD, ILD, and spectral cues-VASI) and on subjective ratings (spatial-SSQ-K). The specific objectives of the study are:

1. To study the effect of musical training on temporal correlate of spatial processing using interaural time difference thresholds (ITD).
2. To study the effect of musical training on intensity correlate of spatial processing using interaural level difference thresholds (ILD).
3. To study the effect of musical training on the composite score (spectral, intensity, and temporal) of spatial processing using virtual acoustic space identification (VASI) test and its perceptual outcomes (SSQ scores).
4. To identify and compare the pattern of spatial errors between musicians and non-musicians.

2 Method

2.1 Participants

Thirty volunteers recruited from Mysore city participated in the current study. The participants were divided into two groups of 15 each based on musical expertise. Group I consisted of participants trained in instrumental music with a minimum experience of 6 months of which (9 males, 6 females; age range 18–30 years, mean: 22.9, SD: 2.87), while Group II comprised of age and gender-matched non-musicians, who had no formal musical training (6 males, 9 females, age range: 18–30 years, mean: 21.6 years; SD: 2.41). The musician group comprised self-identified musicians affiliated with the Raghuleela School of Music in Mysuru, India, and regularly practice music at least 2 h in a day. All the musicians recruited for the study and scored >17 in the questionnaire on music perception ability for Indian classical music [18], indicative of good musical aptitude. All the participants had normal hearing

sensitivity (pure tone average—PTA $\leq$ 15 dB HL). All the participants were free of any otologic complaints, neurologic, cognitive, or other sensory abnormalities. A detailed case history was administered on each participant to ensure the absence of the above-mentioned complaints.

2.2 Informed Consent and Ethical Guidelines

The study followed ethical guidelines of the institutional review board. All participants gave written informed consent mentioning the test details of the study.

2.3 Informed Consent and Ethical Guidelines

Otological history, otoscopy, and tympanometry were performed at the beginning of the enrollment. This was followed by pure tone audiometric test, to determine the hearing thresholds. A battery of three spatial tests and a questionnaire was administered on all the participants in the study. The tests included were interaural time difference (ITD), interaural level difference (ILD), virtual acoustic space identification (VASI) test, and spatial subdomain of SSQ translated in Kannada (SSQ-K). The rationale for the inclusion of spatial tests was their appropriateness in independently manipulating and estimating participant's sensitivity to process any of the three correlates (temporal, spectral, or intensity) of spatial processing. While the test of ITD and ILD provide an understanding of temporal and intensity correlates of spatial processing, VASI test measures the participant's sensitivity on all three spatial processing domains. On other hand, the subjective parameter of spatial acuity is measured using spatial subsection of SSQ-K.

Temporal correlate of spatial acuity—Interaural time difference (ITD) thresholds. ITD thresholds test was administered on all participants in the study using the psychoacoustics toolbox [19] running in MATLAB version 7.10.0 (R2010a) (the Mathworks Inc., Natick). Three white noise bursts (250 ms duration, 5 ms rise and fall time) were presented in each run. Each test stimuli was calibrated to intensity level of 65 dB SPL using sound level meter (B&K, 2270). ITD thresholds were obtained using Sennheiser headphones HD 280 pro (Wedenmark, Germany) connected to the Lenovo laptop.

Stimulus was presented in which three-down one-up transformed staircase procedure. A three-interval forced choice method converging at 79.45% of the psychometric function was used to arrive at ITD thresholds. The participants were instructed to indicate the variable stimulus (the sound which is heard more lateralized to the right ear) among the three stimuli by pressing the number corresponding to that stimuli on the keyboard. The time lag of the variable stimulus was varied depending on the response of the participant. The testing was terminated at 10 reversals, and only the last four reversals were averaged to obtain the converged value of the ITD threshold. The schematic representation of the stimuli is used in ITD test (Fig. 1).

Intensity correlate of spatial acuity—Interaural-level difference (ILD) thresholds. The instrumentation and stimuli used for obtaining ILD thresholds were similar to the one used for obtaining ITD thresholds, except the stimuli in varied interval were louder in right ear instead of time lag (as in ITD). The participants were instructed to indicate the variable stimulus (the sound which is heard louder on the right ear) by pressing the number on the keyboard and the average of 4 reversals were converged to the value the ILD threshold. The schematic representation of the stimuli used in the trails of the study is given in Fig. 2.

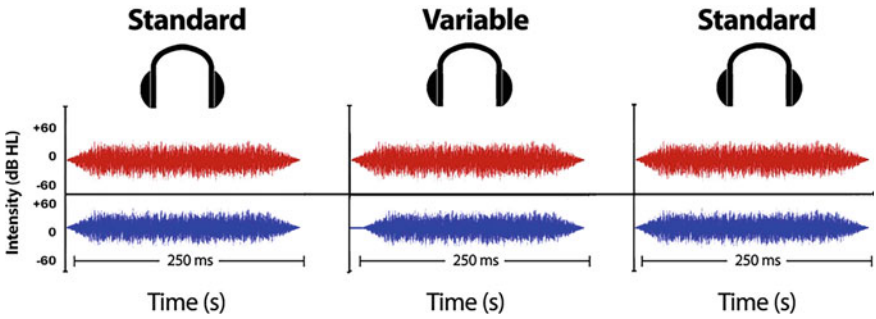


Fig. 1 Schematic representation of the stimuli used in test of interaural time difference (ITD)

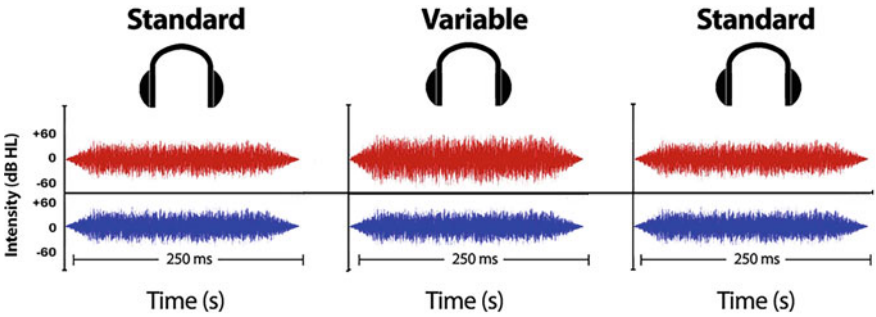


Fig. 2 Schematic representation of the stimuli used in test of interaural-level difference (ILD)

Composite correlate (temporal, intensity, and spectral) of spatial acuity—Virtual auditory space identification test. The stimuli for the VASI test were similar to the one adopted in the study by Nisha and Kumar (2017). The stimuli were created by convolving 250 ms white band noise with HRTF using slab lab [20]. The stimulus produced eight virtual spatial perceptions, i.e., mid-line front (0° azimuth), 45° azimuth toward the right ear (R45), 90° azimuth toward the right ear (R90), 135° azimuth toward the right ear (R135) mid-line back (180° azimuth), and 135° azimuth toward the left ear (L135), 90° azimuth toward the left ear (L90), 45° azimuth toward the left ear (L45). The overall stimuli intensity was 65 dB SPL (calibrated using sound level meter, B&K, 2270). The stimuli were loaded to paradigm software and were presented through a Sennheiser headphones HD 280 pro (Wedenmark, Germany) connected to the Lenovo laptop. Each virtual stimuli was played randomly in all the eight directions (10 times in each direction, collectively the stimuli were presented 80 times). The participants were asked to listen to the stimuli and respond to them by clicking on a dummy head. The test was terminated after the completion of 80 stimuli. The correct locationwise identification scores and the overall accuracy scores are computed. Additionally, a confusion matrix for each participant was also computed by using the modified MATLAB script.

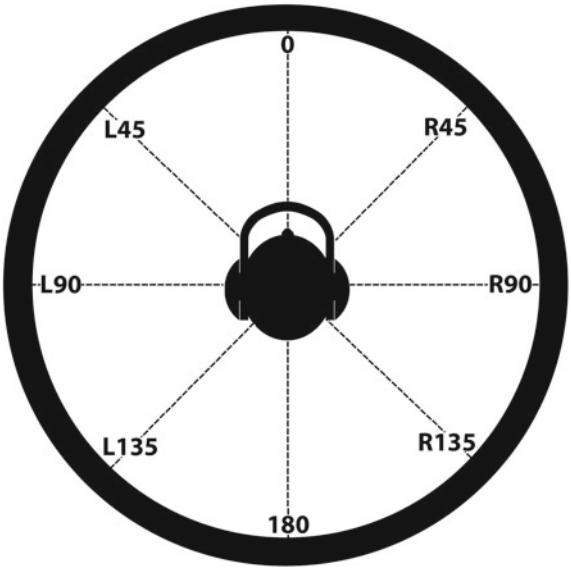


Fig. 3 User interface used for stimulus presentation and response acquisition in the virtual acoustic space identification (VASI) test

Subjective ratings of spatial acuity: Indian version of spatial subsection of SSQ [21]. It is a questionnaire with 14 items rating scale ranging from 1 to 10. The participants are asked to rate all the 14 questions that are applicable in the Indian scenarios. Participants are asked to read the questions given in the questionnaire and assign a score ranging from 1 to 10. The score is given according to the localization ability that the participant can perform the task. Score 1 is given when the participant cannot perform the task at least one time, and a score of 10 is given when the participant can perform the task perfectly all the time. In this manner, the participants are asked to assign a value based upon the ability to do that given task.

2.4 Informed Consent and Ethical Guidelines

The data was subjected to normality test. A multivariate analysis of variance (MANOVA, 2 groups × 4 tests) was carried out to establish main effect of musical training (if any), following which independent t-test was carried out for VASI accuracy. The confusion matrix involving spatial judgment errors was also analyzed for each hemifield and reported separately.

3 Results

Shapiro Wilk's test of normality revealed that the data followed a normal distribution ($p > 0.05$) for all the measures in both musicians and non-musicians. The mean and standard deviation for the two groups are presented in Fig. 4 (Fig. 4a: ITD; Fig. 4b: ILD; Fig. 4c: VASI; Fig. 4d: SSQ), for musicians (blue circles) and non-musicians (red circles). Thresholds were lowest for the musicians for the former two measures (ITD& ILD), while the musicians scored higher than non-musicians for the latter two measures: VASI and SSQ. A multivariate analysis of variance (MANOVA) test with musical training as fixed factors and test scores as dependent factor revealed main effect of musical training [ITD-($F(1,28) = 3.59$, $p > 0.05$; ILD-($F(1,28) = 1.13$, $p > 0.05$; overall VASI scores—$F(1,28) = 11.64$, $p < 0.01$,$\eta_p^2 = 0.29$, SSQ-($F(1,28) = 5.34$, $p < 0.05$, $\eta_p^2 = 0.16$], with musicians showing higher spatial acuity than non-musicians in all tests (Fig. 4).

Detailed analysis of spatial accuracy performance carried out for each virtual location using independent *t*-tests to compare group differences is shown in Table 1, reflective of superior performance of musicians for scores presented in frontal right plane.

Table 2 enumerates the results of confusion matrix analyses documenting the pattern of spatial errors for each virtual location. The spatial confusions are more readily seen in non-musicians compared to musicians for all the virtual locations

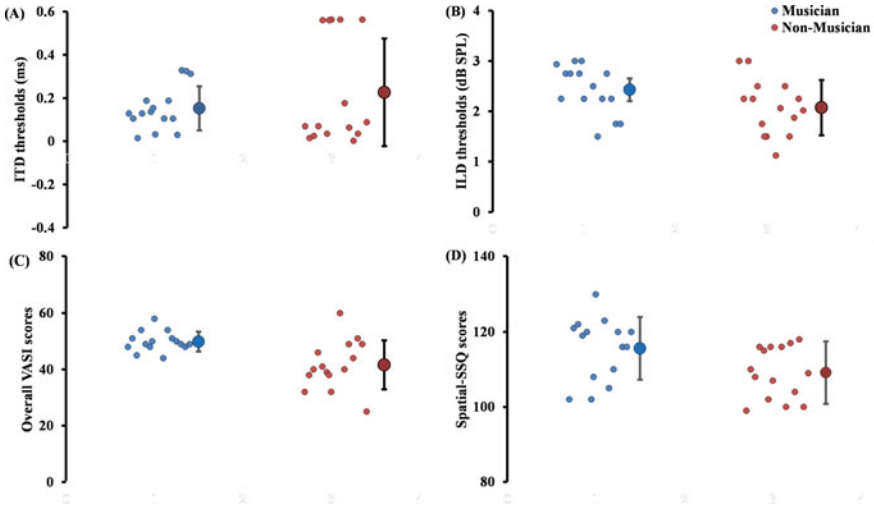


Fig. 4 Individual and mean spatial acuity scores of **a** ITD thresholds **b** ILD thresholds **c** overall VASI scores and **d** spatial SSQ-scores as a function of musical training. The larger circles in each category represent the mean value of the observations for each group, while the vertical error bar indicates standard deviation. The symbols are slightly shifted along the abscissa for better visibility

Table 1 Comparison of performance of musician and non-musician group for each virtual location

Virtual locations	Musician Mean (SD)	Non-musician Mean (SD)	*t* value *t* (28)	Significance (*p*)
R45°	5.86 (2.38)	3.87 (2.06)	2.45	< 0.05
R90°	5.60 (2.74)	5.07 (2.34)	0.57	>0.05
R135°	5.66 (2.29)	4.33 (2.16)	1.64	>0.05
180°	7.00 (2.19)	6.93 (1.98)	0.09	>0.05
L135°	6.73 (1.71)	5.67 (1.76)	1.68	>0.05
L90°	7.60 (1.92)	5.20 (2.36)	1.84	>0.05
L45°	4.93 (2.09)	5.13 (2.03)	0.26	>0.05
0°	7.20 (1.89)	6.07 (2.92)	1.26	>0.05

tested. Complimentary to the above finding, spatial confusions were more in left compared to right hemifield in non-musicians compared to musicians

Table 2 Pattern of spatial errors across virtual locations between musicians and non-musicians

Response VAS location	Group	Target VAS location							
		R45°	R90°	R135°	180°	L135°	L90°	L45°	0°
R45°	Musician	**5.86(2.38)**	1.33(0.98)	2.00(2.17)	0.67(1.05)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.80(0.77)
	Non-musician	**3.87(2.06)**	4.44 (2.38)	3.40(1.92)	0.07(0.26)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.20 (0.41)
R90°	Musician	2.83(3.13)	**5.60 (2.74)**	5.07 (2.34)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.07(0.26)
	Non-musician	3.53(2.90)	**5.07 (2.34)**	2.00(1.31)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.07(0.26)	0.73(1.62)
R135°	Musician	1.07(0.97)	2.00(2.04)	**5.67(2.29)**	1.33(1.11)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.40(0.74)
	Non-musician	2.13(2.20)	0.53(0.74)	**4.33(2.16)**	1.07(1.44)	0.00(0.00)	0.00(0.00)	0.00(0.00)	1.67(0.90)
180°	Musician	0.00(0.00)	0.00(0.00)	0.07(0.26)	**7.00(2.14)**	0.13(0.35)	0.13(0.35)	0.00(0.00)	1.33(1.40)
	Non-musician	0.27(0.46)	0.00(0.00)	0.13(0.35)	**6.93(1.98)**	0.60 (0.83)	0.00(0.00)	0.13(0.35)	1.67(1.45)
L135°	Musician	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.07(0.26)	**6.73(1.71)**	0.93(1.10)	1.53(1.36)	0.13(0.52)
	Non-musician	0.00(0.00)	0.00(0.00)	0.07(0.26)	0.27(0.80)	**5.67(1.76)**	1.07(1.33)	2.07(1.83)	0.67(1.45)
L90°	Musician	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.13(0.52)	1.53(1.55)	**7.60(1.92)**	3.87(1.25)	0.00(0.00)
	Non-musician	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.07(0.26)	1.60(0.74)	**5.20(2.37)**	2.33(2.20)	0.00(0.00)
L45°	Musician	0.00(0.00)	0.13(0.35)	0.00(0.00)	0.07(0.26)	1.40(2.16)	1.33(1.99)	**4.13(1.36)**	0.07(0.26)
	Non-musician	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.20 (0.41)	2.07(1.67)	3.73(1.67)	**5.13(2.03)**	0.00(0.00)
0°	Musician	0.27(0.60)	0.00(0.00)	0.00(0.00)	0.73(0.80)	0.20(0.56)	0.00(0.00)	0.47(0.83)	**7.20(1.90)**
	Non-musician	0.20(0.77)	0.00(0.00)	0.07(0.26)	1.40(1.84)	0.07(0.26)	0.00(0.00)	0.27(0.46)	**6.07(2.91)**

4 Discussion

The present study attempted to understand the benefits of musical training on different dimensions (temporal, intensity, and spectral) of spatial hearing. The results of the current study are discussed under following headings.

4.1 Outcomes of Musical Training on Temporal (ITD Thresholds) and Intensity Correlates (ILD Thresholds) of Spatial Hearing

The results of the temporal and binaural correlate of spatial hearing using ITD and ILD test show that individuals with musical training had significantly lower thresholds compared to individuals who did not receive musical training. This may be attributed to better intensity and temporal resolution abilities in musicians compared to that of non-musicians [6, 7, 22], which in turn can be attributed to large area getting activated in the brain in individuals who underwent musical training [23, 24].

4.2 Outcomes of Musical Training on Composite Score (VASI Test) and Perceptual Scale (SSQ) of Spatial Hearing

VASI is a test of lateralization of sound using virtual auditory sources within the head. The present study reported that musicians have scored significantly lower ($p < 0.05$) overall VASI scores compared to non-musicians (Fig. 4). The higher VASI scores obtained in musicians may be attributed to their superior fine-structure processing skills [25, 26] and memory of pitch [7], both of which improves spatial perception. Analyses of locationwise VASI scores revealed that musicians demonstrated significantly higher accuracy ($p < 0.05$) of virtual source identification in right-frontal plane (R45) compared to non-musicians (Table 1). Reports on right ear superiority of trained musicians for processing of simple auditory stimuli such as melodies or tempo are documented in literature [27]. In addition, the emergence of group differences on SSQ is suggestive of perceptual consequences of heightened spatial awareness in musicians. Sluming et al. [28] demonstrated a strong association between musical ability and visuospatial ability for orchestral musicians, who showed significantly increased activation in Broca's area, and visuospatial network relative to non-musicians. On similar lines, musicians also enjoyed perceptual spatial advantage over non-musicians in reverberant environments, especially in vertical plane [17].

4.3 Comparison of the Pattern of Spatial Errors Between Musicians and Non-musicians

Usually there are confusions arising in localization of sound at different angles even in listeners with normal sensitivity [29]. In general, the present study showed that the degree of confusions is less in individuals receiving musical training compared to the individuals who did not receive any musical training (Table 2) across all the virtual locations tested. This spatial advantage musicians enjoy over non-musicians is much more noticeable in the right hemifield than left, wherein the non-musicians exhibited more spatial confusion in the latter than former. This is in support with several studies on right-ear advantage of musicians over their non-musical counter parts for simple auditory stimuli like tempo or melody [27, 30]. On similar lines, Nelson et al. [31] reported right ear advantage in both musicians and non-musicians for speech-loaded stimuli (dichotic-CV and dichotic-digit free-recall tasks), although no group differences were seen. The lack of group differences in the study can be attributed to the left-hemisphere dominance for processing speech-like stimuli in both the groups, which precluded the right ear advantage otherwise apparent in musicians for simple auditory stimuli.

5 Conclusions

The results of the study provide new evidence for the global (ITD, ILD, VASI, and SSQ scores) advantage of musical training on spatial skills in auditory domain, particularly in the right hemispace. Since the group differences were most significant on VASI test and confusion matrix, we suggest this test can be used as alternative method for the elaborate and expensive loudspeaker-based localization tests for delineating the benefits stemming from musical training on spatial perception. Finally, our results set the stage for future studies on the neural basis exploring the global advantages of musical training on spatial processing. The advantage of musical training in spatial processing can have promising implications on special populations with spatial deficits like auditory neuropathy spectrum disorder (ANSD) and central auditory processing disorder (CAPD), etc., apart from halting age-related deterioration in spatial processing.

References

1. Sares AG, Foster NEV, Allen K, Hyde KL (2018) Pitch and time processing in speech and tones: the effects of musical training and attention. J Speech Lang Hear Res 61:496–509. https://doi.org/10.1044/2017_JSLHR-S-17-0207
2. Marozeau J, Innes-Brown H, Blamey PJ (2013) The effect of timbre and loudness on melody segregation. Music Percept 30:259–274. https://doi.org/10.1525/mp.2012.30.3.259

3. Zendel BR, Alain C (2012) Musicians experience less age-related decline in central auditory processing. Psychol Aging 27:410–417. https://doi.org/10.1037/a0024816
4. Johnson N, Shiju AM, Parmar A, Prabhu P (2020) Evaluation of auditory stream segregation in musicians and nonmusicians. Int Arch Otorhinolaryngol 4:11–20. https://doi.org/10.1055/s-0040-1709116
5. Blauert J (1997) Spatial hearing: the psychophysics of human sound localization. MIT Press, Cambridge. https://doi.org/10.1037/025238
6. Sangamanatha AV, Fernandes J, Bhat J, Srivastava M, Udupa PS (2012) Temporal resolution in individuals with and without musical training. Indian J Speech Hear Assoc 26:27–35
7. Bailey P, Dobinson S (2002) Pitch and loudness memory in musicians and nonmusicians. J Acoust Soc Am 112:2245. https://doi.org/10.1121/1.4778919
8. Donai JJ, Jennings MB (2016) Gaps-in-noise detection and gender identification from noise-vocoded vowel segments: comparing performance of active musicians to non-musicians. J Acoust Soc Am 139:EL128–134. https://doi.org/10.1121/1.4947070
9. Rammsayer T, Altenmüller E (2006) Temporal information processing in musicians and nonmusicians. Music Percept 24:37–48. https://doi.org/10.1525/mp.2006.24.1.37
10. Faßhauer C, Frese A, Evers S (2015) Musical ability is associated with enhanced auditory and visual cognitive processing. BMC Neurosci 16:1–9. https://doi.org/10.1186/s12868-015-0200-4
11. Hansen M, Wallentin M, Vuust P (2013) Working memory and musical competence of musicians and non-musicians. Psychol Music 41:779–793. https://doi.org/10.1177/0305735612452186
12. Helmbold N, Rammsayer T, Altenmüller E (2005) Differences in primary mental abilities between musicians and nonmusicians. J Individ Differ 26:74–85. https://doi.org/10.1027/1614-0001.26.2.74
13. Gaab N, Schlaug G (2003) The effect of musicianship on pitch memory in performance matched groups. NeuroReport 14:342–354. https://doi.org/10.1097/00001756-200312190-00001
14. Horton AM, Monahan MC, Fennell EB (2001) Children's memory scale. Arch Clin Neuropsychol 16:193–198. https://doi.org/10.1093/arclin/16.2.193
15. Wallentin M, Nielsen AH, Friis-Olivarius M, Vuust C, Vuust P (2010) The musical ear test, a new reliable test for measuring musical competence. Learn Individ Differ 20:188–196. https://doi.org/10.1016/j.lindif.2010.02.004
16. Schellenberg EG (2006) Long-term positive associations between music lessons and IQ. J Educ Psychol 98:457–468. https://doi.org/10.1037/0022-0663.98.2.457
17. Kaplanis N, Van Velzen J (2012) Hearing through darkness: a study of perceptual auditory information in real rooms and its effect on space perception. Proc Inst Acoust 34:114–125
18. Neelamegarajan D, Kumar UA, Khyathi G (2017) Development and standardization of questionnaire on music perception ability. Sangeet Galaxy 6(January), 3–13
19. Soranzo A, Grassi M (2014) PSYCHOACOUSTICS: a comprehensive MATLAB toolbox for auditory testing. Front Psychol 5:712. https://doi.org/10.3389/fpsyg.2014.00712
20. Wenzel EM, Miller JD, Abel JS (2000) Sound lab: a real-time, software-based system for the study of spatial hearing. In: 108th AES convention. The Audio Engineering Society, Paris, France, pp 1–27
21. Shetty HN, Palaniappan V, Chambayil SS, Syeda A (2019) Assessment of localization ability—a subjective tool in Kannada Version. J Indian Speech Lang Hear Assoc 33:1–7. https://doi.org/10.4103/jisha.JISHA_2_18
22. Kishon-Rabin L, Amir O, Vexler Y, Zaltz Y (2001) Pitch discrimination: are professional musicians better than non-musicians? J Basic Clin Physiol Pharmacol 12:125–144. https://doi.org/10.1515/JBCPP.2001.12.2.125
23. Ohnishi T, Matsuda H, Asada T, Aruga M, Hirakata M, Nishikawa M, Katoh A, Imabayashi E (2001) Functional anatomy of musical perception in musicians. Cereb Cortex 11:754–760. https://doi.org/10.1093/cercor/11.8.754
24. Gaser C, Schlaug G (2003) Brain structures differ between musicians and non-musicians. J Neurosci 23:9240–9245. https://doi.org/10.1523/jneurosci.23-27-09240.2003

25. Mishra SK, Panda MR, Raj S (2015) Influence of musical training on sensitivity to temporal fine structure. Int. J. Audiol. 54:220–226. https://doi.org/10.3109/14992027.2014.969411
26. Bianchi F, Carney LH, Dau T, Santurette S (2019) Effects of musical training and hearing loss on fundamental frequency discrimination and temporal fine structure processing: psychophysics and modeling. JARO—J Assoc Res Otolaryngol 277:263–277. https://doi.org/10.1007/s10162-018-00710-2
27. Bever TG, Chiarello RJ (2009) Cerebral dominance in musicians and nonmusicians. J Neuropsychiatr 21:94–97. https://doi.org/10.1176/appi.neuropsych.21.1.94
28. Sluming V, Brooks J, Howard M, Downes JJ, Roberts N (2007) Broca's area supports enhanced visuospatial cognition in orchestral musicians. J Neurosci 27:3799–3806. https://doi.org/10.1523/JNEUROSCI.0147-07.2007
29. Makous JC, Middlebrooks JC (1990) Two-dimensional sound localization by human listeners. J Acoust Soc Am 87:2188–2200. https://doi.org/10.1146/annurev.ps.42.020191.001031
30. Majak J, Śliwinska-Kowalska M (2016) Does musical training enhance auditory processing test performance in normal hearing adults? Otorynolaryngologia 15:165–172
31. Nelson MD, Wilson RH, Kornhass S (2003) Performance of musicians and nonmusicians on dichotic chords, dichotic CVs, and dichotic digits. J Am Acad Audiol 14:536–549. https://doi.org/10.3766/jaaa.14.10.2

Effect of Transformation of Vowel Duration on Speech Intelligibility in Children with Hearing Loss

S. V. Narasimhan and N. P. Nataraja

Abstract The study aimed was to document the effects of the transformation of vowel duration in the utterances of the children with hearing loss on speech intelligibility and acceptability. The present study was carried out under two experiments. In experiment 1, the words uttered by the participants of both groups were acoustically analysed, and the vowel duration was extracted from the target vowels. Descriptive and inferential statistics were carried out to the significant differences in vowel duration between the children with hearing loss and normal hearing. In experiment 2, the vowel duration of all the target vowels in the utterances of the children with hearing loss was transformed (i.e. the duration of the vowel was modified) to approximate the normal values (as in the speech of the children with normal hearing), and the perceptual evaluation, to determine the intelligibility and acceptability of speech, for the transformed utterances, unaltered utterances of the participants with hearing loss and utterances of the children with normal hearing was carried out. The results of experiment 1 showed that there were significant differences in the vowel duration between the participants of groups. The results of experiment 2 showed that there were significant improvements in intelligibility and acceptability of speech of the children with hearing loss after the transformation of vowel duration.

Keywords Transformation of speech · Vowel duration · Hearing loss · Hearing impairment · Analysis by synthesis

1 Introduction

Several researchers have reported the effect of hearing loss on the acquisition of speech [1–4]. A child with hearing loss (CHL) encounters severe communication problems in two ways. Normal speech is unintelligible to the CHL, and due to the lack of auditory feedback of his/her speech production, the child encounters considerable difficulty in learning to speak correctly [5].

S. V. Narasimhan (✉) · N. P. Nataraja
JSS Institute of Speech & Hearing, Mysore, Karnataka, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_23

The aural-oral communication skills in CHL have long been of concern to educators, speech-language pathologists and audiologists as social, educational and career developments of these children depend on their communication abilities. The ultimate goal of aural-oral communication has been to attain communication skills as in the case of CNH [6]. Attempts to teach speech to the CHL span over several decades. Yet, the competence of an CHL to make himself understood remains to be poor [7]. Only about 20% of the speech of the CHL may be understood by the "person on the street" [8].

Research to determine the factors responsible for poor intelligibility of speech in CHL has been going on. Speech intelligibility has been studied with relation to suprasegmental and segmental errors in the speech of CHL. According to Paatsch and Blamey, both suprasegmental and segmental errors cause poor intelligibility of speech in CHL [9]. Thus, both segmental and suprasegmental features of the speech of the CHL have been described by researchers. Therefore, to develop more effective speech training procedures for the CHL, it is necessary to study the deviation in their speech from that of the CNH, and hence, the analysis of acoustic parameters in the speech of the CHL becomes essential.

Vowel duration is one such important temporal parameter that has been studied by various investigators, under different conditions. Duration from the onset of the vowel to the vowel offset is referred to as vowel duration [10] measured in terms of milliseconds [11].

The actual duration of any particular vowel has been found to depend on its height, the position in the word, accentual or tonal characteristics, rate of speech, grammatical complexity, adjacent segments, word length and the physical and psychological state of the individual. The theoretical motivation guiding studies of temporal aspects of speech have been found to quantify the effects of each of these factors and to postulate hypothesis concerning the temporal organisation of speech which has been considered to explain the variability of speech segments. Mechanisms that underlie the temporal organisation of speech have been reported to be very complex and not understood [10].

Vowel duration is one of the powerful factors to determine both phonetic and phonemic quality of the vowels. There have been several studies that have investigated vowel duration in the utterances of the CHL and CNH and have reported significantly longer vowel durations in the utterances of the CHL compared to the speech of the CNH [12–20]. Thus, the literature review shows that the vowel duration plays an essential role in the speech intelligibility and vowel duration to be longer in the utterances of the CHL and CNH.

Development of technology has permitted studies using computer synthesis to determine the role of different factors on intelligibility of speech. Several investigators have tried to eliminate these confounding variables by using computer processing techniques. In this regard, the use of transformation of the utterances of the CHL and CNH children offers several advantages over the traditional speech correction studies. Transformation of the acoustic and temporal parameters offers control over the parameters of speech and permits manipulation of specific acoustic and/or temporal parameter to the desired manner, without modifying other parameters. This

helps to determine the effect of transformation of specific acoustic parameter and combinations of these parameters on the speech intelligibility. This also permits to determine the possible relationship between the modifications of the parameters and the intelligibility by controlling the other variables. Therefore, the current investigation aimed at studying the effects of the transformation of the duration of vowel in the speech of the CHL on the intelligibility and acceptability of speech.

2 Method

The present study was carried out under two experiments. In experiment 1, the words uttered by the participants of both groups were acoustically analysed, and the vowel duration was extracted from the target vowels. Statistical analysis was performed to calculate the mean, standard deviation and the significant differences in vowel duration between the subjects of both groups. In experiment 2, the vowel duration of all the target vowels in the utterances of the CHL was transformed (i.e. the duration of the vowel was modified) to approximate the normal values (as in the utterances of CNH), and the perceptual evaluation, to determine the intelligibility and acceptability of speech, for the transformed utterances, unaltered utterances of CHL and utterances of CNH was carried out.

2.1 Experiment 1

Participants. Group 1 included ten male CHL in the age range of 8–10 years, and group 2 included ten male CNH matched in terms of age. The number of participants was determined based on the statistical power analyses using G-Power (ver. 3) [21] using the mean and standard deviation (SD) values of the parameters of the speech of CHL reported by earlier investigators [11, 22]. The power of the test was assigned a value of 0.8 and the α error probability of 0.05.

All the participants of group 1 had congenital sensorineural hearing loss bilaterally. The pure tone average was larger than 70 dBHL in the better ear [23] as per the audiometric testing. They were attending speech and language therapy and using hearing aid regularly at least for the past three years. All the participants of group 2 had normal hearing sensitivity in both ears [23] as per the audiometric testing. Speech and language skills were age-appropriate. Participants of both groups were native speakers of Kannada and were able to read Kannada words. No participant had any history of neurological, psychological or other problems. All the participants were from middle socioeconomic status as per socioeconomic status scale [24]. Informed consents indicating the willing to take part in the investigation were obtained from the parents of each participant.

Materials. Eighty simple bisyllabic meaningful Kannada words (CVCV) (without geminate clusters) with voiced and unvoiced bilabial (/p/ and /b/), dental (/t/ and /d/), velar (/k/ and /g/) and retroflex (/ṭ/ and /ḍ/) stop consonants in initial and/or medial position and vowels /a/, /i/, /u/, /e/ and /o/ were selected from Kannada textbooks meant for children studying in the first, second and third standard. In order to select the words that were most familiar to the children aged 8–10 years, three primary school teachers, who were native speakers of Kannada, were given the list of words and were instructed to rate each word based on the familiarity to the children of 8–10 years. They were asked to rate the word as "not familiar", if the word was used by the children in the class less than 50% of time, as "familiar", if the word was used approximately 50–75% of time by the children in the class and "most familiar" if the word was used approximately more than 75% of time by the children in the class. Twenty words which were rated as "most familiar" by all the teachers were selected.

Data collection procedure. A sound-treated room used for audiometry which met the ANSI specifications prescribed for hearing evaluation [23] was used for data collection. The room was free from distraction and well lit. In order to present the printed words, each word was typed on cards (6″ × 4″ size) to elicit the speech from the participants. Out of twenty cards, sixteen cards were used to collect the stimuli and four cards were used for trial or practice sessions. The speech sample was recorded from one participant at a time. The participant was made to sit across the table, opposite to the investigator. The CNH was instructed in Kannada by the investigator, at a comfortable loudness level through the auditory modality, to read the word written on the card. Similar instructions were given to the CHL in the auditory and visual modality (using gestures). Each CHL wore the hearing aids suitable for his/her hearing loss during the evaluation. One card was shown to the participant at a time and was asked to read. As the participant uttered the word, it was recorded using a microphone (Sony F-V120 Dynamic, unidirectional microphone with a frequency response of 60–12,000 Hz) that was placed approximately at a distance of eight centimetres from the mouth of the participant. A Dell laptop with Intel Core i7 processor loaded with Praat software (version 6.0.56) [25] was used to record and analyse the verbal responses of the participants. A sampling frequency of 44.1 kHz was used to record the utterances, using mono recording at 16 bits quantization. Using this procedure, all the sixteen cards were shown, one by one and each utterance along with the three repetitions of the utterances, as a response to the card presented, and were recorded. Thus, each participant had uttered 16 words, three times each, yielding 48 utterances which were recorded using the procedure explained earlier. Adopting this procedure, the speech samples of each participant of both groups were recorded and stored on the hard disc.

Acoustic analysis of vowel duration. The vowel duration of was measured for the vowels /a/, /i/, /e/, /u/, /a:/, /o:/ and /e:/ using Praat software in the utterances of the participants of both groups. Vowel duration was measured in milliseconds by placing the cursor at the beginning of regular glottal pulses to the final periodic

pulses excluding the portion of formant transition associated with the vowel on the spectrogram and waveform. Thus, the vowel duration values extracted from all the words uttered by the participants of both groups were tabulated and next statistically analysed.

2.2 Experiment 2

Transformation of vowel duration. Transformation involved the modification of the vowel duration in the speech of the CHL to approximate the normal vowel duration values as in the speech of the CNH. Transformation of vowel duration in the speech of the CHL was carried out by reducing the vowel duration as the results of the statistical analysis showed that the vowel duration values were significantly higher in the speech of the CHL. The vowel duration was modified using the Praat vocal toolkit. Thus, the vowel duration values of both medial and final vowels in all the sixteen words uttered by all the CHL were modified to approximate the normal mean vowel duration values, and results were saved as separate .wav files on the hard disk of the computer for perceptual analysis.

Perceptual analysis. Three qualified female speech-language pathologists and three female naïve listeners were selected to carry out the perceptual evaluation of the unaltered and the transformed utterances of the CHL. A sound-treated room used for audiometry which met the ANSI specifications prescribed for hearing evaluation [23] was used. The room was free from distraction and well lit. The perceptual analysis was carried out individually by all the six judges. Each judge was seated comfortably, and the loudspeakers (Zebronic SW1000 2.1 multimedia speakers with frequency response of 40 Hz–20 kHz) connected to the laptop were placed in front of the judge. Each judge was presented the original unaltered, and the vowel duration was transformed utterances of CHL through the loudspeaker and was instructed to carefully pay attention to each word played and to rate the intelligibility of the words on a three-point rating scale where “1” referred to “unintelligible”, “2” referred to “partially intelligible” and “3” referred to “highly intelligible speech”. Speech acceptability was also rated by the judges on a two-point rating scale where “0” refers to “not acceptable”, and “1” refers to “acceptable speech”. The ratings were tabulated and were statistically analysed.

3 Results and Discussion

3.1 Experiment 1

The average values of vowel durations of three repetitions of each target vowel were determined in milliseconds and are presented in Table 1. The vowel /o:/ had shown

Table 1 Mean (in ms), standard deviation (SD) and range of vowel duration of the vowels in the words uttered by the participants of both groups

Target vowel	Group 1		Group 2	
	Mean	Standard Deviation	Mean	Standard Deviation
a	246.69	86.45	192.12	31.65
i	276.23	126.59	221.12	37.32
e	246.13	92.24	174.55	30.92
u	261.42	108.90	236.33	35.36
a:	217.22	66.14	196.34	32.31
o:	276.91	121.37	235.64	43.56
e:	206.42	58.87	298.58	46.75

the highest mean, and the vowel /e:/ had shown the lowest mean values in group 1. The vowels /i/ and /e:/ had the highest and lowest standard deviation values. The vowel /e:/ and vowel /e/ had shown the highest and lowest mean and standard deviation values of vowel durations, respectively, in the utterances of participants of group 2. The mean vowel durations of group 1 were higher than group 2 except for the vowel/e:/. Group 1 showed higher variability in the vowel durations than that of group 2. The range of the vowel duration was also higher in group 1 compared to that of group 2.

The results of the independent samples t-test carried out at 0.05 level of significance, and comparing the vowel duration between the participants of both groups is displayed in Table 2. It was observed from Table 2 that there were significant differences in the vowel duration between group 1 and group 2 in all the vowels, except for /i/, /u/, and /o:/. Therefore, it was concluded that the CHL showed significantly longer duration for the vowels compared to the CNH.

Shukla had also reported that the duration of vowels was longer and showed greater variation, in CHL compared to the CNH. The study had also reported that both groups of subjects had not shown any consistent variation in the duration of the vowels

Table 2 Results of the independent sample *t*-test comparing vowel durations in the utterances of the participants of both groups

Target vowel	t value	df	*p*-value
a	−2.70	18	0.01*
i	−1.91	10.77	0.08
e	−2.55	13.13	0.02*
u	−0.35	18	0.72
a:	−2.98	12.07	0.01*
o:	−1.12	18	0.27
e:	−2.80	9.85	0.01*

*Significance at the level of 0.05

Table 3 Percentage of intelligibility and acceptability of the unaltered words and the words with the transformation of the vowel duration in the utterances of the CHL

Listeners	Intelligibility of speech			Acceptability of speech		
	HLU (%)	HLT (%)	PI (%)	HLU (%)	HLT (%)	PI (%)
Listener 1	69	89	20	82	93	11
Listener 2	70	89	19	82	94	12
Listener 3	70	86	16	82	89	7
Listener 4	67	84	17	85	93	8
Listener 5	66	82	16	85	88	3
Listener 6	65	85	20	81	93	12
Average	68	86	18	83	92	9

HLU—Unaltered utterances by the CHL; HLT—Transformed utterances by the CHL; PI—the percentage of improvement in the transformed utterances

depending on the preceding consonants [17]. Many Indian studies have strongly reported that the CHL had significantly longer vowel duration values compared to that of the CNH [11–14, 16, 18, 26, 27]. Sony had also reported that the vowel duration was substantially longer, and the SD values were also higher in the speech of the CHL compared to that of CNH [27]. As the present study also showed that the vowel duration in CHL was longer and variability was also higher compared to the CNH, it can be stated that the present study results are in consonance with the findings of the earlier studies. Thus, the CHL had significantly longer vowel duration values compared to the CNH, as per the results of the present study.

3.2 Experiment 2

Table 3 shows the percentage of intelligibility and acceptability of speech rated by the listeners for the unaltered words and the words with the transformation (modification of the vowel duration) in the utterances of the CHL. The results showed that the listeners had reported higher intelligibility scores for the vowel duration transformed words compared to the unaltered words. The average intelligibility scores of the unaltered words and transformed words were 68 and 86%, respectively, with an improvement of 18% in intelligibility after transformation. The average acceptability scores of the unaltered words were 83%, and transformed words were 92% leading to an improvement of 9% in acceptability of speech following transformation of vowel duration. The significance of differences in the acceptability and intelligibility ratings of the transformed and the unaltered words in the speech of the CHL was determined using Mann Whitney U test. The results of the test showed that there were significant differences in both intelligibility ($/z/ = -13.95$, $p < 0.05$) and acceptability ($/z/ = -5.34$, $p < 0.05$) between the transformed and the unaltered words.

Therefore, the results of the present study revealed that the transformation of the vowel duration caused a significant improvement in intelligibility and acceptability in the speech of the CHL. Duration has been found as the feature to an accurate expression of contrasts in vowels and other segments [28–30]. Therefore, the modifications of temporal aspects of speech in the utterances of the CHL have been found to improve the intelligibility of speech [31].

Earlier researchers have also carried out the transformation of temporal, spectral and prosodic parameters in the speech of the CHL and have reported that the transformation of the temporal parameters resulted in maximum improvement in the intelligibility of speech [11]. Sheela had studied the effect of computer correction of some of the temporal aspects in the speech of the CHL on speech intelligibility and found that the maximum improvement in intelligibility had been observed when the vowel duration was corrected. Improvement was also noticed that when pause alone was corrected. But when the combination of vowel duration and pause was corrected, the intelligibility scores were less [18]. Jagadish had also reported that the corrections of vowel duration with the elimination of pauses resulted in improvement in the intelligibility of speech of the CHL [32]. Rasitha had also studied the speech of Malayalam speaking CHL in the age range of five to nine years and the effect of synthesis of speech of these subjects on the intelligibility and speech and had found that the intelligibility had improved when the vowel durations were corrected, and speech intelligibility was reduced when the inter-syllabic pauses were corrected along with the vowel duration and pauses [33]. The results of the present study are also in agreement with these studies as the present study also reported highest improvements in speech intelligibility and acceptability of speech when vowel duration was transformed in the speech of the CHL.

It was also noted that the transformation resulted in improvement of 18% in the intelligibility and 9% in the acceptability of speech in the utterances of CHL. This might be due to the reason that all the CHL had already had shown better scores of intelligibility and acceptability of speech (in their unaltered utterances), and therefore, the unaltered and transformed utterances might not have caused considerable perceptual differences. Thus, the lower percentage of improvements in intelligibility and acceptability of speech might have been noticed after the transformation of the speech of the CHL in the present study compared to the previous studies [11, 32]. However, the results also showed that the transformation of vowel duration in the speech of CHL did not result in a 100% improvement in speech intelligibility or speech acceptability scores. Therefore, it may be inferred that other acoustic parameters may be contributing to the improvement in intelligibility and acceptability of speech.

4 Conclusion

The present study aimed to investigate the effects of the transformation of vowel duration in the utterances of the CHL on speech intelligibility and acceptability. The

present study was carried out under two experiments. In experiment 1, the words uttered by the participants of both groups were acoustically analysed, and the vowel duration was extracted from the target vowels. Descriptive and inferential statistics were carried out to determine the significant differences in vowel duration between the CHL and CNH. The results of experiment 1 showed that there were significant differences in the vowel duration between the participants of groups. In experiment 2, the vowel duration of all the target vowels in the utterances of the CHL was transformed (i.e. the duration of the vowel was modified) to approximate the normal values (as in the speech of the CNH), and the perceptual evaluation, to determine the intelligibility and acceptability of speech, for the transformed utterances, unaltered utterances of the CHL and utterances of the CNH was carried out. The results of experiment 2 showed that there were significant improvements in intelligibility and acceptability of speech of the CHL after the transformation of vowel duration. Further studies can focus on effects of transformation of other temporal parameters of speech in CHL to broaden the knowledge on the contribution of various parameters in improvement of speech intelligibility.

References

1. Eilers RE, Oller DK (1994) Infant vocalizations and the early diagnosis of severe hearing impairment. J Pediatr 124:199–203
2. Ertmer D, Mellon J (2001) Beginning to talk at 20 months: early vocal development in a young cochlear implant recipient. J Speech, Lang Hear Res 44:192–206
3. Moeller MP, Hoover B, Putman C, Arbataitis K, Bohnenkamp G, Peterson B, Lewis D, Estee S, Pittman A, Stelmachowicz P (2007) Vocalizations of infants with hearing loss compared with infants with normal hearing: Part II—transition to words. Ear Hear 28:628–642
4. Rvachew S, Slawinski EB, Williams M, Green CL (1999) The impact of early onset otitis media on babbling and early language development. J Acoust Soc Am 105:467–475
5. Levitt H, Smith RC (1972) Errors of articulation in the speech of profoundly hearing-impaired children. J Acoust Soc Am 51
6. Osberger MJ, Mc Garr NS (1982) Speech production characteristics of the hearing impaired. Speech Lang 8:221–283
7. Ling D (1976) Speech and the hearing impaired child: theory and practice. Alexander Graham Bell Association for the Deaf, Washington
8. Gold T (1980) Speech production in hearing-impaired children. J Commun Disord 13:397–418
9. Paatsch LE, Blamey PJ (2001) Effects of articulation training on the production of trained and untrained phonemes in conversations and formal tests. J Deaf Stud Deaf Educ 6:32–42
10. Gopal NK (1986) Acoustic analysis of speech in normal adults
11. Nataraja NP, Savithri SR, Sreedevi N, Sangeetha N (1999) Tranformation of the speech of the hearing impaired
12. Priya (1998) Acoustic analysis of speech of Malayalam speaking hearing impaired children
13. Rahul (1997) Acoustic transformation of speech of hearing impaired
14. Nataraja NP, Savithri SR, Sreedevi N, Sangeetha (1998) Analysis of the speech of the hearing impaired. In: International conference on computational linguistics speech & document processing, Culcutta
15. Christy G (2016) Analysis of vowels in the speech of hearing impaired children
16. Shahid HB (2008) Acoustic Analysis of Speech in Telugu speaking hearing impaired children
17. Shukla RS (1987) Objective measurements of the speech of the hearing impaired

18. Sheela K (1988) Analysis and synthesis of hearing impaired speech
19. Vandam M, Ide-Helvie D, Moeller MP (2011) Point vowel duration in children with hearing aids and cochlear implants at 4 and 5 years of age. Clin Linguist Phonetics 25:689–704
20. Poonam (1998) Acoustic analysis of speech of Punjabi speaking hearing impaired children
21. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191
22. Chithra NV (2011) Acoustic analysis of speech of Malayalam speaking hearing impaired children
23. ANSI S3.21:2004 (2009): Methods for manual pure-tone threshold audiometry. Bioacoustics
24. Venkatesan (2009) Readapted from 1997 Version NIMH socio economic status scale. Secunderabad
25. Boersma P, Weenink D (2019) Praat: doing phonetics by computer [Computer program]
26. Jayaprakash E (1998) Temporal and acoustic analysis of speech of Kannada speaking hearing-impaired children
27. Sony G (2015) Acoustic analysis of speech of Malayalam speaking children with hearing impairment
28. Cooper, Delattre, Liberman, Borst, Gerstman (1952) Some experiments on the perception of synthetic speech sound. J Acoust Soc Am 24:597
29. Lisker L, Abramson AS (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word J Int Linguist Assoc 20:384–422
30. Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects
31. Metz DE, Schiavetti N, Samar VJ, Sitler RW (1990) Acoustic dimensions of hearing-impaired speakers' intelligibility: segmental and suprasegmental characteristics. J Speech Hear Res
32. Jagadish MS (1989) Synthesis of speech of the hearing impaired
33. Rasitha (1994) Analysis and synthesis of speech of hearing impaired

The Influence of Music on Image Making: An Exploration of Intermediality Between Music Interpretation and Figurative Representation

Pinaki Gayen, Junmoni Borgohain, and Priyadarshi Patnaik

Abstract Music evokes emotional responses; it also evokes visual imagery that is closely linked with the emotions evoked. To explore this inter-medial relationship, an experimental study was conducted with 70 students aged between 18 and 22 years. Depending upon audience response ratings, two contrast emotion (happy and sad-calm) evoking Indian classical instrumental (*sitar*) *rāga* music clips were selected previously. Both the music clips were played separately in two different classrooms and the participants were asked to respond to these clips through depiction of figurative paintings. The participants were also asked to provide a title for each of the paintings with some description about these depictions. Contents of the paintings were analyzed in terms of their themes, usage of lines and colors. Findings revealed that students depicted sad and calm emotional contents with dull-dark colors while listening to sad music and happy music made them depict happy emotional contents with light and bright colors. From semiotic analysis of the depicted paintings, we could clearly identify inter-medial relationships between certain visual features and certain musical elements like tempo. Findings revealed that emotions play a key role as a bridge in connecting these two mediums.

Keywords Emotion · Figurative painting · Inter-medial relationship · Music

1 Introduction

Music and visual arts share certain correspondences. While one is temporal and the other spatial, both elicit emotions, share common aesthetic criteria such as balance, harmony and structure, and often occur together, as in multimedia. Listening to music often conjures images in the listener's mind that can induce emotions in the process [1]. Neuroscience researchers have observed and demonstrated this phenomenon as the involuntary musical imagery (INMI), which appears spontaneously and without

P. Gayen (✉) · J. Borgohain · P. Patnaik
Department of Humanities and Social Sciences, Indian Institute of Technology Kharagpur, Kharagpur, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_24

conscious control of the listener [2–7]. Music can induce episodic memories triggering reminiscing of previous events. But while experiencing emotions, a listener can also be influenced by images and events not related directly to episodic memory. Such images may be influenced or shaped by similarities in structure [1] or other cultural associations.

Music can stimulate vivid imagery in our mind and certain musical characteristics, like slow tempo, repetitive patterns, predictability in rhythmic, melodic and harmonic elements, are more effective in creating those imageries [8]. Tempo has been identified as one of the significant musical features and has a major impact on emotional experience [9]. A number of studies have found that listeners have a tendency to associate faster tempo and major modes with happiness, and slower tempo and minor modes with sadness [10–14]. Researchers have observed a relationship between musical tempo and color where color associations change if the tempo of a music clip is manipulated to a faster or slower one [15].

Cross-modal matches between music and colors used in paintings are fundamentally mediated by emotional associations [16]. Music can change the evaluation of representational and abstract painting [17]. On the other hand, music evaluation is more affected by the presence of different types of paintings [18]. Correspondingly, perceptual similarities in the aesthetic structure of music and paintings have been identified by researchers [19].

The review of literature suggests that most studies have been done on western music. Moreover, the role of cultural associations has also not been explored. Further, earlier studies have explored the association of music and images primarily through the mode of recognition. This study uses excerpts from *Hindustani classical music* as stimuli to study their influence on figurative image making. The study also examines the inter-medial relationship between music and evoked images through a semantically driven approach.

2 Research Questions

1. What representational strategies are used by the participants while responding to the *Hindustani classical music clips* through figurative visuals?
2. Do emotions play a key role in the way that the two mediums are interrelated?

3 Method

3.1 Participants

Seventy undergraduate students of the Indian Institute of Technology Kharagpur participated in the study. Their age range varied from 18 to 22 years ($M = 40$, $F = 30$). None of the participants had any prior training either in music or in visual arts.

Table 1 Music clips used in the experiment

Music	Emotion	Tempo	Time (s)
Komal Rishav Asavari	Sad-calm	Slow	00.00.30
Jaunpuri	Happy	Fast	00.00.30

3.2 Stimuli

Two *Hindustani classical rāga music* clips played in the same instrument (*sitar*) at two different tempos—slow and fast—were used in the experiment. They were rated previously by 70 participants in a self-report method. The choice was based on seven emotion labels where they had to select the emotion that best represented the clip. The emotion labels were—happy, sad, anger, fear, romantic, wonder and calm, based on the *Navarasa* theory of emotions elucidated in the *Bharata's Natyashastra.* Disgust was omitted as it is a rare musical emotion [20]. The clips used in this experiment got significant responses on happiness and sadness. Their details are given in Table 1.

3.3 Materials

The materials which were supplied to the participants consisted of white papers ($11'' \times 16''$), multiple color crayons from the brand Camlin, 2B, 4B, 6B pencils and erasers.

3.4 Procedure

The experiment was conducted in two different classrooms with 35 participants in each group. Before starting the experiment, the participants were seated comfortably and briefed about the experiment. They were told that they will listen to a music clip in a repetitive manner. The music might evoke certain visuals in their mind constantly and at the same time the participants were asked to depict the evoked visual imageries in the form of figurative painting. After they were assured of the instructions, they were provided the materials. One of the groups listened to *Komal Rishav Asavari* (slow tempo music clip) and the other to *Jaunpuri* (fast tempo music clip). The time period given to complete the task was 1 hour.

Fig. 1 Representation of slow tempo music

Fig. 2 Representation of fast tempo music

3.5 Analytical Strategy

The visual data was analyzed using the inter-rater reliability method. Three competent visual artists who were involved in the evaluation of both sets of data adopted a two-staged approach. At first, the images were classified as positive and negative on the basis of the overall content of the visuals (subject of the image, qualitative content and use of color). In the second stage, the emotions portrayed in the visuals and colors used were analyzed through visual semiotic methods (iconic, symbolic and indexical). A few sample images of the figurative paintings representing the slow tempo and fast tempo music clips are given in Fig. 1 and Fig. 2 respectively.

4 Results

The content analysis of figurative paintings revealed that participants depicted sad, peaceful and devotional events to represent the music clip in slow tempo. The fast tempo music clip was represented through joyful, peaceful and romantic image contents. In case of slow tempo music, the distinctive colors used to express sadness were mainly gray, brown, deep blue and black. On the other hand, peaceful emotion

was depicted using colors such as orange, light blue and light yellow (Table 2). The devotional aspect was expressed with colors like light yellow and orange. Visual contents used to denote sadness in sad clip were human being in negative moods and scene of death. Visual contents used to denote other emotions in sad clip were tree without leaves, landscape, human being worshiping God, etc.

The distinctive colors used by the participants to express joyful emotion for the fast tempo music clip were mainly light yellow and light green. Colors used to express peaceful emotion in fast tempo clip were mainly light blue and light brown. For the representation of romance in fast tempo music clip, the participants used mainly light yellow and red color. Visual content used to denote emotions in the fast tempo music were musical instrument and dance (Table 3).

Semiotic analysis was applied to find out if some aspects of musical features or responses influence visual representations of the two *rāga* music clips. Peircean semiotic constructs were used to analyze data. Images and elements where certain structural features of music were emulated were treated as iconic (resemblance), while components that were elicited by association or connection were treated as indexical. Where cultural arbitrary components came in, based on convention, they were treated as symbolic [21].

The results revealed that most of the participants have used iconic representations in terms of representing the *rāga* music clips in figurative paintings. Iconic elements were the representations such as musician playing musical instruments, dancing figure beside music instrument, musical notes, and hand movement on music instruments (Table 4).

For slow tempo music clip, the content of the paintings depicted sad events experienced in a human life. For instance, images of death, human being in negative mood (crying) lonely person, natural hazards and tree without leaf were represented. Peaceful paintings were associated with the images of landscapes, human being in positive moods (walking, seating in natural environment, etc.), musical instruments and musicians, etc. The devotional paintings were associated with human being in meditation, worshiping God and religious festivals. For fast tempo music, the joyful paintings were associated with musical instruments, musician, music concert, dance, festivals, marriage, birthday, landscapes, etc. (Table 4).

5 Discussion

The music clips of the study with fast tempo (happy) and slow tempo (sad and peace) features were consistent with an experimental study where it has been found that happiness ratings were elevated for fast tempo and major-key stimuli, sadness ratings were elevated for slow tempo and minor-key stimuli [13]. But the present study found some significant lesser known aspects while these *Hindustani classical raga* music clips (fast and slow tempo) were represented through figurative paintings. The results of the present study revealed that the fast tempo music elicited not only happy emotion but also other emotions such as peaceful and romantic, whereas the slow

Table 2 Colors used to depict images and emotions conveyed by the music clips

Slow Tempo			Fast Tempo		
Emotion	Color	(%)	Emotion	Color	(%)
Sadness	Gray	26.24	Joyful	Light yellow	19.57
	Brown	14.01		Light green	14.13
	Deep blue	12.17		Light blue	11.96
	Black	10.55		Orange	9.78
	Yellow ochre	10.21		Brown	9.78
	Light blue	10.02		Deep blue	8.70
	Red	8.63		Red	6.52
	Light yellow	6.44		Deep green	6.52
	Orange	4.37		Black	4.35
Peaceful	Orange	15.49		Pink	3.26
	Light blue	14.08		Violet	3.26
	Light yellow	14.08		Gray	1.09
	Light green	11.27		Sap green	1.09
	Brown	8.45	Peaceful	Light blue	21.74
	Deep blue	8.45		Light brown	21.74
	Red	8.45		Light yellow	17.39
	Sap green	5.63		Light green	13.04
	Gray	4.23		Orange	8.70
	Violet	4.23		Deep blue	8.70
	Black	4.23		Pink	4.35
	Deep green	1.41		Gray	4.35
Devotional	Light yellow	16.67	Romance	Light yellow	23.53
	Orange	14.58		Red	17.65
	Light blue	12.50		Light green	11.76
	Brown	10.42		Orange	11.76
	Light green	8.33		Pink	11.76
	Black	8.33		Violet	11.76
	Deep green	6.25		Deep blue	5.88
	Deep blue	6.25		Deep green	5.88
	Gray	6.25			
	Red	6.25			
	Violet	2.08			
	Pink	2.08			

Table 3 Semiotic analysis of visual representation of music clips

Music clip	Emotions	Indexical (%)	Iconic (%)	Symbolic (%)
KomalRishavAsavari	Negative	9.09	33.33	10.6
	Positive	3.05	36.36	7.57
Jaunpuri	Positive	13.64	71.21	15.15

Table 4 Thematic analysis of visual representations of music clips

Komal Rishav Asavari (slow tempo)		*Jaunpuri (fast tempo)*	
Visual contents	(%)	Visual contents	(%)
(In positive moods)		(In Positive Moods)	
Landscape	16.47	Musical instrument	24.19
Human being	14.12	Dance	20.97
Musical instrument	3.53	Landscape	20.97
(In negative moods)		Human being	19.35
Human Being	25.88		
Death	11.76		
Tree without leaves	9.41		
Hill	2.35		
Animal	4.71		
Sun	5.88		
Hut	2.35		
Cloud	1.18		
Water drop	1.18		
Musical instrument	1.18		

tempo music clip evoked emotions such as sadness and peace, which is consistent with the study of Gabrielsson [9].

We found a great degree of similarity in terms of emotions expressed through music clips and emotions represented through figurative drawings. This finding is consistent with a study where the perceptual similarities between classical music clip and figurative images have been identified [19]. In case of slow tempo music, the colors used in the paintings were mostly dark colors (gray, brown, deep blue and black). On the other hand, in case of fast tempo music, participants used light colors (light yellow, light green, light blue) to depict joyful, peaceful and romantic image contents. These color preferences for representing fast tempo music through figurative paintings are consistent with the study of Tsang and Schloss [15] where they found that music in the major mode and at faster tempo was associated with lighter colors. In similar lines, it was observed that brighter colors such as yellow, red, green and blue were associated with happy songs and gray to the sad songs [22].

Therefore, the association between the two representative modes, music and images drives us inferring that there exists a common link between the two mediums.

Depending upon their forms, their correlations with other things, or their involvement in conventional relationships, three different kinds of representative relationships can be derived between two things—they are iconic, indexical, and symbolic relationships [21]. The semiotic analysis reveals that iconic representation in terms of decoding the *rāga* music clips into figurative painting is high. In some cases, we also found indexical and symbolic representations of certain musical elements like slow and fast tempo in these figurative paintings. Thus, semiotic indicators can provide insights into the nature of existing inter-medial relationship between a music piece and a painting.

6 Conclusion

The study provides insights into the relationship between Indian classical music and figurative paintings. One of the strongest connecting links appears to be emotions which play the role of a link between music and visual images (figurative painting). Thus, sadness in music finds expression in images that can depict sadness either through figurations and socio-cultural contexts or through colors. While Juslin and Vastjfall [1] explore why episodic or anecdotal memory is not pertinent in music evoked images, we find that semiotic resemblances play a key role in not only evocation of images, but their representation as well. Thus, musical features and structures such as rhythm, tempo, musical instruments, themselves contribute extensively in what is painted. On the other hand, indexical components like rejoicing in marriage, deliberate closing of eyes while praying among the others play a key role in the act of visual representation.

In future, it would be worthwhile to go for qualitative analysis and detailed exploration of the visual as well as appended textual responses to music in order to identify how the mind works to evoke images in response to music. Similarly, how the evoked imageries related to a particular music piece can be combined with the music at hand, such as visuals in movies, music videos, or cultural festivals would also be another rich area to explore. The role of gender in musical and visual interpretations can also be meaningfully explored.

Acknowledgements The authors are grateful to the students of the Indian Institute of Technology Kharagpur for their participation in this study.

References

1. Juslin PN, Västfjäll D (2008) Emotional responses to music: the need to consider underlying mechanisms. Behav Brain Sci 31(5):559–575
2. Farrugia N, Jakubowski K, Cusack R, Stewart L (2015) Tunes stuck in your brain: the frequency and affective evaluation of involuntary musical imagery correlate with cortical structure. Conscious Cogn 35:66–77
3. Halpern AR (2001) Cerebral substrates of musical imagery. Ann N Y Acad Sci 930(1):179–192
4. Herholz SC, Halpern AR, Zatorre RJ (2012) Neuronal correlates of perception, imagery, and memory for familiar tunes. J Cogn Neurosci 24(6):1382–1397
5. Hubbard TL (2010) Auditory imagery: empirical findings. Psychol Bull 136(2):302
6. Liikkanen LA (2008) Music in every mind: Commonality of involuntary musical imagery. In: Proceedings of the 10th international conference on music perception and cognition, Aug 2008. ICMPC, Sapporo, Japan, pp 408–412
7. Zatorre RJ, Halpern AR, Perry DW, Meyer E, Evans AC (1996) Hearing in the mind's ear: a PET investigation of musical imagery and perception. J Cogn Neurosci 8(1):29–46
8. McKinney CH, Tims FC (1995) Differential effects of selected classical music on the imagery of high versus low imagers: two studies. J Music Ther 32(1):22–45
9. Gabrielsson A (2009) The relationship between musical structure and perceived expression. In: Oxford handbook of music psychology
10. Crowder RG (1984) Perception of the major/minor distinction: I. Historical and theoretical foundations. Psychomusicol: J Res Music Cogn 4(1–2):3
11. Dalla Bella S, Peretz I, Rousseau L, Gosselin N (2001) A developmental study of the affective value of tempo and mode in music. Cognition 80(3):B1–B10
12. Gagnon L, Peretz I (2003) Mode and tempo relative contributions to "happy-sad" judgements in equitone melodies. Cogn Emot 17(1):25–40
13. Hunter PG, Schellenberg EG, Schimmack U (2010) Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions. Psychol Aesthet Creat Arts 4(1):47
14. Webster GD, Weir CG (2005) Emotional responses to music: interactive effects of mode, texture, and tempo. Motiv Emot 29(1):19–39
15. Tsang T, Schloss KB (2010) Associations between color and music are mediated by emotion and influenced by tempo. Yale Rev Undergrad Res Psychol 82
16. Palmer SE, Schloss KB, Xu Z, Prado-León LR (2013) Music–color associations are mediated by emotion. Proc Natl Acad Sci 110(22):8836–8841
17. Limbert WM, Polzella DJ (1998) Effects of music on the perception of paintings. Empir Stud Arts 16(1):33–39
18. Parrott AC (1982) Effects of paintings and music, both alone and in combination, on emotional judgments. Percept Mot Skills 54(2):635–641
19. Actis-Grosso R, Lega C, Zani A, Daneyko O, Cattaneo Z, Zavagno D (2017) Can music be figurative? Exploring the possibility of cross-modal similarities between music and visual arts. Psihologija 50(3):285–306
20. Zentner M, Eerola T (2010) Rhythmic engagement with music in infancy. Proc Natl Acad Sci 107(13):5768–5773
21. Deacon TW (1998) The symbolic species: the co-evolution of language and the brain. WW Norton & Company
22. Barbiere JM, Vidal A, Zellner DA (2007) The color of music: correspondence through emotion. Empir Stud Arts 25(2):193–208

Comparison of Neural Network Architectures for Speech Emotion Recognition

Shubham Patil and Saaket Agashe

Abstract Human speech encapsulates a large amount of information. The ability of modern machine learning algorithms to understand and recognize speech is that of vital importance. Recognizing emotions in speech can be of great importance in a large number of fields from psychology to human–computer interaction. Speech emotion recognition (SER) has been a challenge in signal processing for a long time; however, recent advances in deep learning have taken us a lot closer to reliable SER. We propose a set of deep learning models for the classification of emotions in speech. We use both the audio as well as text as features and compare them head to head. We also make a comparison between models within audio and text paradigms. The audio models presented include convolutional neural networks, gated recurrent unit networks, and a combination of both of these. The text models on the other hand include a long short-term memory network and a gated recurrent unit network, both of which are appended with an attention layer. We test our models on the IEMOCAP dataset and present the result of our findings. We show how each model generalizes in the task and how they learn from the given features.

Keywords Deep learning · Long short-term networks · Convolutional neural networks · Speech emotion recognition · IEMOCAP · Spectrogram · Gated recurrent units · Embeddings

1 Introduction

Deep learning and machine learning have proven the ability to recognize, translate, and understand human speech. One powerful endeavor toward further development in artificial intelligence is the ability to capture and understand human speech. Speech emotion recognition finds great use in fields like human-computer interaction where machines not only have to understand human emotions but also respond to them.

S. Patil (✉) · S. Agashe
Department of Electrical Engineering, Veermata Jijabai Technological Institute, Matunga, Mumbai, India
e-mail: shubhamppatil30@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_25

Consequently, speech emotion recognition is a curious and important problem in signal processing.

We aim to analyze the various nuances between models and features for the optimal classification of emotions in the IEMOCAP dataset [4]. While the dataset primarily contains eight different emotion classes, we focus our attention toward only four of them, particularly "Neutral," "Angry," "Sad," and "Happy." This particular decision is made to stay at pace with contemporary literature. We present a comparative study of different models and how well they perform on the dataset in question. The first step toward reliable SER is the extraction of optimal salient features from audio and text data. Researchers have explored an array of features ranging from prosodic features, cepstral coefficients like mel-frequency cepstral MFCC, and delta MFCC as well as spectrograms [15–17]. While MFCCs have conventionally been the most used features, the advent of convolutional neural networks is making the visual spectrograms more and more useful. In our audio-based speech recognition model, we propose the use of mel scaled spectrograms as the primary features.

The use of word embeddings for text features has found great popularity. Sentiment classification on the text has been explored in great detail, in examples like the classification of human sentiments through Twitter tweets [26]. However, multiclass sentiment classification, as in the case of the problem at hand remains a much more difficult problem. While pre-trained embeddings have found great use, we use a set of trainable custom embeddings in our model which gives better results for us than pre-trained embeddings. There is a multitude of algorithms and models proposed in the field of deep learning including convolutional neural network, recurrent models like gated recurrent unit, and long short-term memory networks as well as the novel attention framework. It is important to try and understand the optimal models for the task of speech emotion recognition. In this paper, we present a comparative analysis of the various models applied to both audio and text data of the IEMOCAP dataset.

2 Related Work

Speech emotion recognition (SER) is a branch of research that focuses on inferring emotions from speech signals. In this section, we will review some previous works in this field of research. This field of research is not new and has been studied for quite some time in contemporary literature. Recent studies [6] have focused on distinguishing appropriate acoustic features such as fundamental frequency (pitch), the intensity of utterance, bandwidth, and duration for speech analysis [7]. Some have combined information from all the three modalities to extract emotion. Feature level fusion of audio and textual modalities for emotion recognition was also done [9, 10]. Term frequency-inverse document frequency (TFIDF), a textual features extraction technique was used by [8] for speech emotion recognition. Traditional approaches for identifying emotion from speech [11, 12] used hidden Markov models (HMMs) [13]. The introduction of deep learning leads the researchers to propose various neural network-based architectures to improve the performance in this domain. In the

primary study, deep neural networks (DNNs) demonstrated its efficiency in speech emotion recognition by extracting high-level features from speech [14]. With the advancement of deep learning methods and easy availability of powerful computing platforms, more complex neural network-based architectures were developed such as convolutional neural network (CNN), LSTMs, etc. Researchers have experimented with these models using features like spectrograms, mel-frequency cepstral coefficients (MFCCs), and low-level descriptors (LLDs) [15–17]. A combination of these neural network models [4, 18] was applied to the IEMOCAP dataset, and these models recorded the best performance.

3 Dataset

In this work, we use the IEMOCAP [4] released in 2008, one of the widely used databases on speech emotion recognition. Researchers at the University of Southern California (USC) have compiled a database containing recorded sessions of conversations from ten speakers. That amounts to nearly 12 h of audio-visual information along with transcriptions. It is annotated with eight categorical emotion labels, namely anger, happiness, sadness, neutral, surprise, fear, frustration, and excitement. The corpus is divided into five sessions containing conversations between two actors each. Furthermore, no two actors overlap between these sessions. In addition to that, each session is further split into multiple utterances for each session. This was done with the help of the start timestamp and end timestamp provided for the transcribed sentences. This results in a total of approximately 10K audio files which are then used to extract features.

3.1 Data Preprocessing

The IEMOCAP database contains scripted and improvised dialogs in the form of videos. We only used transcription and audio data. In our case, we have decided to work with speech utterances from four emotion categories, i.e., happy, sad, angry, and neutral. We opted for merging examples from“excited” and “happy” classes as “happy” was underrepresented, and two emotions share a similarity. Furthermore, we discarded examples that were classified as “others,” as the corresponding examples were labeled ambiguous for humans. We used the data from four sessions to train the model, while the remaining session was used for validation and testing, thus achieving fivefold cross-validation across the dataset in a way that does not have any speakers overlapping in the train and test datasets (Table 1).

Audio: There are 5584 audio recordings in the IEMOCAP having varying durations. Consequently, we first plotted the durations to obtain the median value and searched

Table 1 Number of samples for each emotion class

Emotions	Samples
Happy	1689
Sad	1084
Neutral	1708
Angry	1103

for the best results around this value. The ideal duration of 5 s was established by checking around the median value of duration distribution which was 4 s.

Text: Raw data may contain a lot of useless information which may lead to misleading results. To overcome this issue, we have removed the special symbols such as punctuations and abbreviations and then normalized the transcriptions.

3.2 Feature Extraction

3.2.1 Audio Features

We used the mel-spectrograms as features for the audio model. The primary motivation was that the mel-scale was how humans distinguish frequency information, i.e., better distinction is possible between lower frequencies than higher ones. We also determined experimentally that Mel-spectrograms outperformed linear spectrograms for our models. We used the Librosa library for audio processing and used their in-built mel-spectrogram function. We first compute the short-time Fourier transform of the audio recording. We follow this with a fast Fourier transform on every window. We then scale the spectrogram to the mel-scale and follow it by converting the spectrogram to a dB representation. The parameters used in spectrograms are shown in Table 2. The spectrograms were plotted and visually inspected to get an idea of their differences based on the emotions as shown in Fig. 1.

Table 2 Spectrogram parameters

Parameter	Value
Window type	Hanning window
FFT Window Length	2048
Hop length	512
mels	128
Top dB value	80

3.2.2 Text Features

The notion behind deep learning is that the neural network studies representations of the features, rather than expecting the researcher to design them. There exist some latent semantic attributes that the network is adept in learning. That is why in our case, we just let the word embeddings be parameters in our model, and then be updated during training. We have split each utterance into a list of words and then define an index for each word when using embeddings. We can access the embeddings for each word from the $|V| \times D$ matrix such that the word assigned index acts as the key, i.e., index i corresponds to i-th row of the matrix containing its word embedding, where the inputs are $|V|$ dimensional, V is our vocabulary, and D is the dimension of the embeddings. Padding is done to generate inputs of the same length to train our recurrent neural network.

4 Neural Network Architectures

4.1 Audio Models

4.1.1 Convolutional Neural Network

We use a four-layer CNN network as shown in Fig. 2. Our CNN uses a deep stride architecture similar to the one proposed in [3]. The input image is a 128×157 dimension spectrogram. The first layer has 32 filters, padding of 1, and a kernel size of (5×5). We use a stride length of 2 for downsampling the feature vector. The consecutive layers have 64, 128, and again 128 filters with the same kernel size and padding, as well as the same stride length to downsample the image size. Each convolution layer is followed by the ReLU activation and a batch normalization layer. We use the batch normalization layer after the ReLU [5]. The output of CNN is passed to a dense layer of 1024 units, followed by another dense layer of 512 units. Both of these layers also use ReLU activation. Finally, the output is obtained from a four-layered (corresponding to Neutral, Angry, Sad, and Happy) dense layer with a softmax activation.

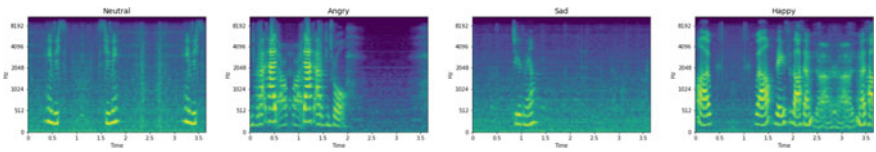


Fig. 1 Spectrograms of different emotions

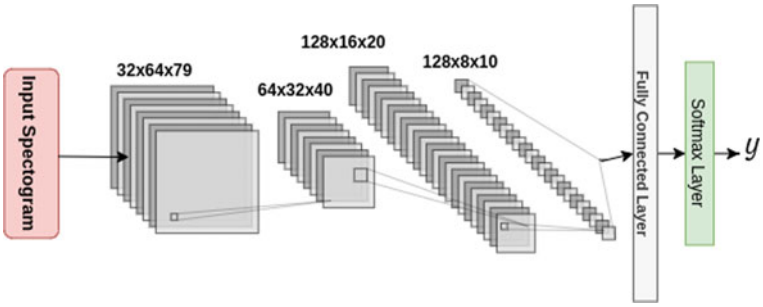


Fig. 2 Convolutional Neural Network Architecture

4.1.2 Bidirectional Gated Recurrent Unit (GRU)

Gated recurrent unit [24] is nothing but a modified version of existing long short-term memory (LSTM). One major issue faced by the RNNs is the vanishing gradient problem which impedes the learning of long data sequences. To resolve this problem, gated recurrent units resort to update gate and reset gate. These two vectors can be trained to decide what information is relevant to the prediction and hence to be passed through to the output or remove irrelevant information. We use a two-layer GRU network with the spectrogram frequency information as to its features and time information as the sequence. The GRU has 1024 hidden units, two layers, and it is bidirectional. There are three subsequent dense layers of size 1024 units, 512 units, and four units. The final layer is passed through a softmax to output the emotion.

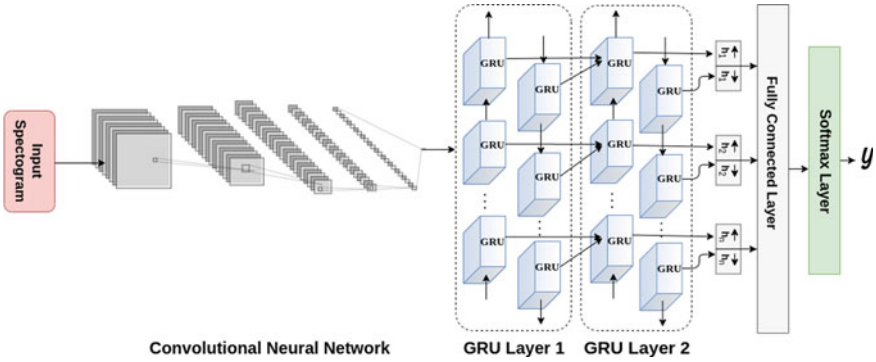


Fig. 3 Convolutional Neural Network + 2 Layer Bidirectional GRUs

4.1.3 Convolutional + Gated Recurrent Unit Neural Network

In this model, we pass the output of CNN from the previously mentioned model through a two-layer gated recurrent unit as shown in Fig. 3. The output of CNN is of the form "Height x Width x Filters." The width of the output is used as the time step for GRU, while the filters and height are stacked together as features for the GRU. The GRU makes use of 1024 hidden units and is bidirectional. As in the previous cases, there are three subsequent dense layers of size 1024 units, 512 units, and four units, and the final layer is passed through a softmax to output the emotion.

4.2 Text Models

See Fig. 4.

4.2.1 Embedding Layer

To extract the meaningful information from the utterances, each utterance is firstly represented as a dense vector of word embeddings. Denote u as transcription of the utterance with n-words and each word is mapping to a global vector [19], then we have:

$$u = [| \overrightarrow{e_1} || \overrightarrow{e_2} || \overrightarrow{e_3} | \ldots\ldots\ldots | \overrightarrow{e_n} |] \tag{1}$$

where vector e_i represents the vector of ith word with a dimension of D. The vectors of word embeddings are concatenated together to maintain the order of words in an utterance. Consequently, it can overcome the shortfall of bag-of-words techniques. For our methods, the embedding dimension of the embedding layer is $D = 256$.

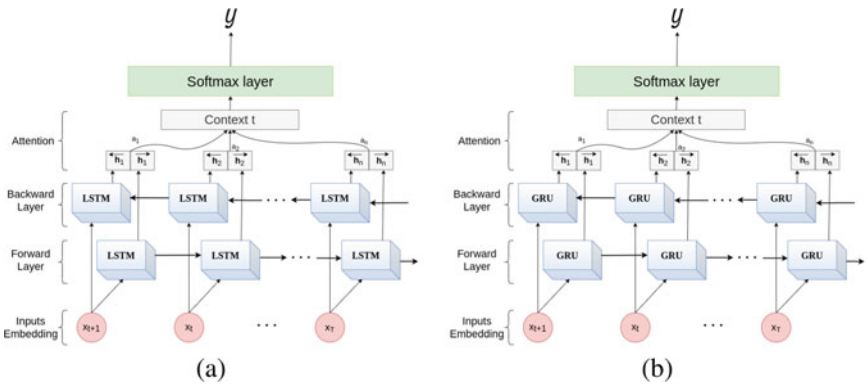


Fig. 4 **a** Bidirectional LSTM + Attention, **b** Bidirectional GRU + Attention

During the creation of the embedding layer, the tensors are initialized randomly. It is only after you train the layer is when this similarity between similar words appears. We also analyzed using the embedding layer seeded with the GloVe word embedding weights. In this case, we do not update the learned word weights during the training, and hence we set the trainable attribute to be false (see Fig. 4a, b).

4.2.2 Bidirectional Long Short-Term Memory (LSTM) Layer

Traditional neural networks do not seem to have the ability to retain information. However, a branch of neural networks called recurrent neural networks (RNN) allows previous outputs to be used as inputs, allowing them to retain information. In emotion recognition task, we model the utterances using RNNs, especially we examine bidirectional LSTM + Attention [20, 21]. Bidirectional LSTMs are an extension of standard LSTM [22]. According to current research [23], there are advantages of using such networks to consolidate contextual information in the classification process. In our case, the model utilizes an attention mechanism with bidirectional LSTM to record the most significant semantic information in a sentence (see Fig. 4a). We have also experimented using a bidirectional GRU + Attention, and it seems to increase the efficacy of the algorithm. After the embedding layer, the sequence of word vectors is fed into a bidirectional LSTM or bidirectional GRU to obtain another representation for word vectors, i.e., h_{ti}. Bidirectional LSTM/GRU will generate a sequence of forward hidden states $\overrightarrow{h_i}$ and a backward one $\overleftarrow{h_i}$ for each input sequence. A simple concatenation will represent the encoder state. The aim is to combine both the preceding and following words in the annotation of one word.

$$h_i = [\overrightarrow{h_i}^T; \overleftarrow{h_i}^T], \text{where } i = 1, 2, 3, \ldots n \tag{2}$$

4.2.3 Attention Layer

The attention mechanism is the technique focusing on the essential elements to correlate words in one sentence. While representing an utterance, the contribution of all the words in the utterance is not uniform, by leveraging the word attention mechanism [25] we can ascertain the unique significance of the words on the emotion of the utterance and generate a context vector combining the features from each time step into a sentence-level feature vector. Specifically,

$$u_{ti} = \tanh(W_w h_{ti} + b_w) \tag{3}$$

$$\alpha_{ti} = \frac{exp(u_{ti}^T, u_w)}{\sum_{k=1}^{n} exp(u_{tk}^T, u_w)} \tag{4}$$

$$s_t = \sum_{i=1}^{n} \alpha_{ti} h_{ti} \tag{5}$$

where W_w and b_w represent weights and biases, t represents t^{th} utterance, i represents i^{th} word in the utterance, and n is the number of annotations in an utterance.

Primarily, we feed the word annotation h_{ti} (word annotation of i-th word in t-th utterance) through a single-layer MLP to get u_{ti} as a hidden representation of h_{ti} (see Fig. 4a, b). The next step is calculating the importance of the words as the similarity of u_{ti} with a word-level context vector u_w. The same step is accompanied by getting a normalized importance weight α_{ti} (the weight of i-th word in the t-th utterance) through a softmax function. The context vector is initialized randomly and learned in conjunction during the training process. Greater α_{ti} indicates that the i-th word is significant for emotion representation. Finally, the last step is representing the sentence vector s_t as a weighted sum of the word annotations. These vectors are further passed through a softmax layer to output the emotions.

5 Experiments

5.1 Evaluation Metrics

The IEMOCAP dataset that we have utilized is highly imbalanced in terms of classes, with the neutral emotion being the most prevalent class. We, thus, adopt both unweighted accuracy and weighted accuracy as our metrics for classification.

- Weighted accuracy—the fraction of correct predictions in a class averaged across all the classes.
- Unweighted accuracy—the overall average of accuracy across all the utterances.

We further take a weighted loss function, to restrict the models from overfitting to one class.

5.2 Results

5.2.1 Audio

The audio samples were passed through a five-fold validation, where one session was kept as the test sample, while the model was trained on four of the remaining sessions. The generated spectrograms were normalized to a mean of 0 and a standard deviation of 1. The models were trained using the Adam [27] optimizer for 20 epochs with a learning rate of 2×10^{-5}. A weighted loss function was added to compensate for the class imbalance and avoid overfitting to one class and ensure better generalization. We

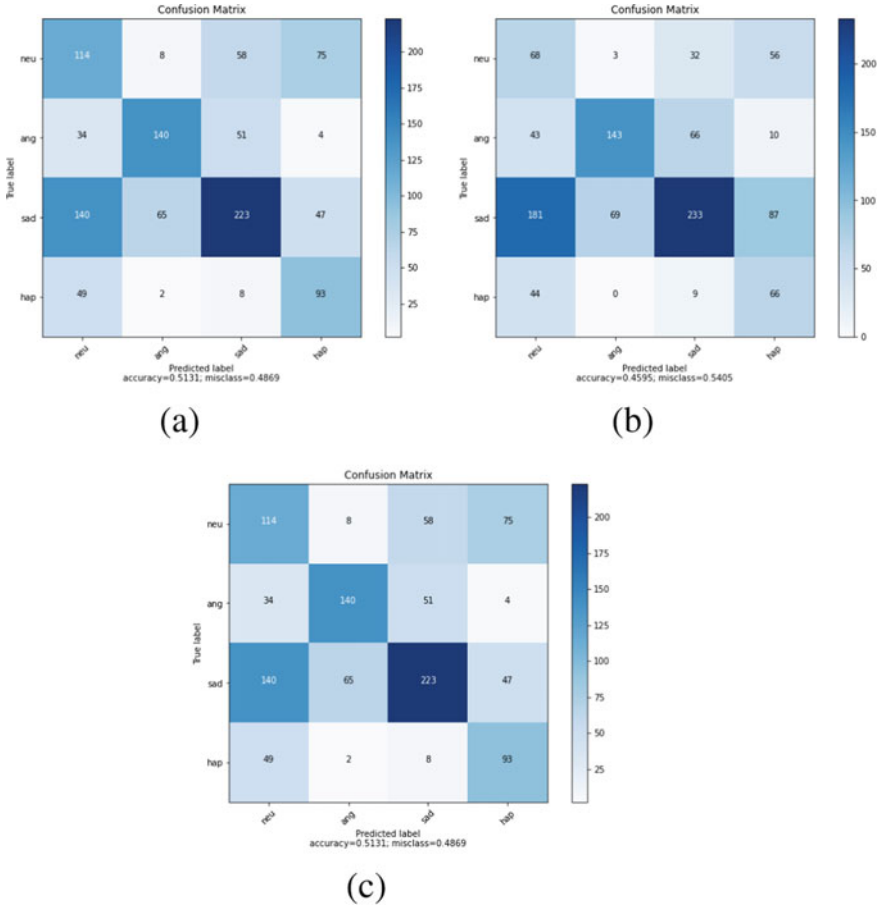


Fig. 5 **a** CNN, **b** GRU, **c** CNN + GRU

also used a learning rate of decay of gamma = 0.1 every 10 epochs. The experimental results for the average of the five-fold cross-validation are summarized in Table 3. We also include the confusion matrix corresponding to the four emotions in Fig. 5.

Table 3 Comparative results of audio model

Model	Unweighted accuracy (%)	Weighted accuracy (%)
CNN	51.36	53.28
CNN-GRU	51.25	53.56
GRU	45.98	49.51

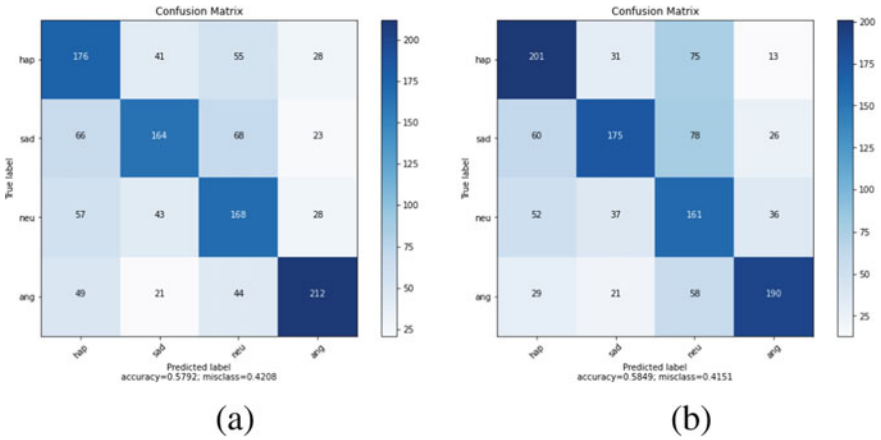


Fig. 6 **a** Bidirectional LSTM + Attention, **b** Bidirectional GRU + Attention

Table 4 Comparative results of text models

Model	Unweighted accuracy(%)	Weighted accuracy(%)
Bidirectional LSTM + Att	57.92	58.63
Bidirectional GRU + Att	58.49	60.00

5.2.2 Text

For the processing of text transcriptions, we first encoded the sentences. This was followed by padding the sentences to the max length of the sentence. We also created a dataset-specific vocabulary and used custom embeddings instead of using pre-trained ones, as they resulted in better performance. The training loop for textual classification uses the Adam optimizer and was trained for 50 epochs at a steady learning rate of 0.001. We also used a weighted loss function as in the case of audio models. As seen in the text classification, there is an attention layer corresponding to both of the models. The input to the attention layer is moderated by reshaping, and the output is then passed to the fully connected layer to classify the emotion of the utterance. The experimental results for the average of the five-fold cross-validation are summarized in Table 4, and the confusion matrix is shown in Fig. 6.

6 Conclusion and Future Work

In this paper, we examined three audio models and two text models, elaborated their details, and presented the results obtained for each model. We observed that the pure recurrent neural network does not perform well enough for the task of extracting

features from a spectrogram and is outperformed by the CNN. We also noted that the IEMOCAP dataset has a class imbalance, and without a weighted loss function, the network overfits to only one class and misclassifies others. Furthermore, the use of custom text embeddings in our text models proved better than using pre-trained GloVe embeddings.

Given these results, we can see that an appropriate combination of audio and text features can work toward improving the results. Furthermore, the use of transformers is a good direction for further research and may prove beneficial.

References

1. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2017) Hierarchical attention networks for document classification. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1480–1489
2. Yoon S, Byun S, Jung K (2018) Multimodal speech emotion recognition using audio and text. In: IEEE Spoken Language Technology Workshop (SLT). Athens, Greece 2018, pp 112–118. https://doi.org/10.1109/SLT.2018.8639583
3. Mustaqeem, Kwon S (2019) A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors 20(1):183
4. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: interactive emotional dyadic motion capture database. J Language Resources Eval 42(4):335–359 December
5. Mishkin D, Sergievskiy N, Matas J (2016) Systematic evaluation of convolution neural network advances on the Imagenet. Comput Vision Image Understanding. ISSN-1077-3142
6. Datcu D, Rothkrantz L (2008) Semantic audio-visual data fusion for automatic emotion recognition. In: Euromedia'2008. https://doi.org/10.1002/9781118910566.ch16
7. Wollmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, Morency LP (2013) Youtube movie reviews: Sentiment analysis in an audio-visual context. IEEE Intelligent Syst 28(3):46–53
8. Sahu G, Cheriton DR. Multimodal speech emotion recognition and ambiguity resolution.*arXiv preprint* arXiv:1904.06022
9. Metallinou A, Lee S, Narayanan S (2008) Audio-visual emotion recognition using gaussian mixture models for face and voice. In: Tenth IEEE International Symposium on ISM 2008. IEEE, pp 250–257
10. Eyben F, Wöllmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R (2010a) On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. J Multimodal User Interfaces 3(1–2):7–19
11. Nogueiras A, Moreno A, Bonafonte A, Mariño JB (2001) Speech emotion recognition using hidden Markov models. In: Seventh European conference on speech communication and technology
12. Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003 Proceedings. (ICASSP'03), vol. 2, pp. II- 1, IEEE
13. Rabiner LR, Juang B-H (1986) An introduction to hidden Markov models. IEEE ASSP magazine 3(1):4–16
14. Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In: Fifteenth annual conference of the International Speech Communication Association

15. Bertero D, Fung P (2017) A first look into a convolutional neural network for speech emotion detection. In : 2017 IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5115–5119
16. Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 International conference on Platform Technology and Service (PlatCon). IEEE, pp 1–5
17. Aldeneh Z, Provost EM (2017) Using regional saliency for speech emotion recognition. In: 2017 IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 2741–2745
18. Satt A, Rozenberg S, Hoory R (2017) Efficient emotion recognition from speech using deep learning on spectrograms. In: Proc. Interspeech 2017, pp 1089–1093
19. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781
20. Graves A, Mohamed A-R, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 6645–6649
21. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw 18(5–6):602–610
22. Sepp H, Jurgen S (1997) Lstm can solve hard long time lag problems. In: Advances in neural information processing systems, pp 473–479
23. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: The 54th Annual Meeting of the Association for Computational Linguistics, pp 207–213
24. Cho K, van Merrienboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)
25. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2017) Hierarchical attention networks for document classification. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 1480–1489
26. Zhou Q, Wu H (2018) NLP at IEST 2018: BiLSTM-attention and LSTM-Attention via soft voting in emotion classification. WASSA@EMNLP
27. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

A Comprehensive Review on Analysis Methods, Software, and Application of fMRI for Classification of Alzheimer's Disease

Chintan R. Varnagar and Hitesh B. Shah

Abstract Neurodegeneration refers to the progressive loss and damage to the structure and/or function of neuronal cell, leading to death of neurons. It is one of the most feared associates of the increased longevity. Alzheimer's disease (AD) is a neurodegenerative disorder, whose etiology and path physiology is not known to the fullest. The capacity of recognizing the first sign of disease has enormous socioeconomic benefit. Applications of various noninvasive MR imaging techniques are widely explored by researchers for this purpose. Here we reviewed (1) the recognized pathophysiology of Alzheimer's disease and its categories; (2) data analysis methods applied on functional magnetic resonance imagining (fMRI) modality and summarized them with respect to statistical power, method, approach, its application(s), advantage and limitation of them; (3) Reviewed often used fMRI data processing tool box/software with respect to different aspects and features it offers; (4) a critical review on application of artificial intelligence (AI) techniques w.r.t objective(s) achieved, approach followed, methods applied, materials used, modalities considered, dataset used, process followed, accuracy achieved/claimed and results reported by various researchers to classify AD subject from fMRI data and suggest multimodality-based approach for computer-aided detection (CAD) system, to improve accuracy of classification.

Keywords Alzheimer's diseases · Functional MRI · AD biomarkers · Computer-aided diagnostic · Classification · fMRI data processing software

C. R. Varnagar (✉)
Computer Engineering Department, Government Engineering College—Rajkot, Gujarat Technological University, Ahmedabad, Gujarat, India
e-mail: chintan2287@gmail.com

H. B. Shah
Electronics and Communication Engineering Department, G. H. Patel College of Engineering & Technology, Vallabh Vidyanagar, Gujarat, India
e-mail: hiteshshah@gcet.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_26

1 Introduction

Alzheimer's disease (AD) is referred as progressive and irrevocable neurodegenerative disorder in which neuron gradually loses its capacity to function (degenerates over time) and ultimately leads to abnormal death of neuronal cell. It adversely affects brain regions which are responsible for cognitive functions such as memory, language, thinking and problem solving, attention, planning, and time and space orientation. In AD patient, these cognitive functions exhibit decline as the disease progresses, ultimately loses ability to perform very basic task/activity independently, which makes the patients' lives demanding. It is the most feared associates of increased longevity, globally as per the WHO and other government departments/agencies reports [1], it is the sixth primary reason of death in the USA in older adults, nearly 1 out of 7 people is affected, $234 billion is spent for medicine and care of the patient, and therefore, it is insistent public health issue and financial problem for society.

The exact and specific cause for AD its underlying pathology, pathophysiology, progression, role of genetics, etc., is not fully understood and known yet [1]. However, at the core or as the root cause, there is misfolding of brain protein. These get deposited into neuronal cell, due to which the cell fails to (1) function normally (2) unleash a set of toxic events (3) loses connection with each other, also, (4) multiple neuron cell dies, resulting in (5) structural brain atrophy (shrinkage in volume) [1–3]. Researchers have intensively studied role of two proteins with respect to pathological changes seen in AD patients:

1. Extracellular neuritic amyloids plaques—*Beta-amyloid protein (Aβ)*, a leftover or residual fragment of a superior protein group, when clusters together, has a toxic consequence on neurons and disrupts inter-cell communication.
2. Tangles—*Tau protein* is responsible for neuron's interior support, and it is a transport system which carries nutrients. In AD, tau protein changes shapes and organizes themselves into structure called neurofibrillary tangles, i.e., deposition of intra-neuronal aggregates of cytoskeleton elements, which disrupts transport system [1, 2].
3. This disruptive and progressive process begins years before the first symptom is felt or evident, mostly in the region of the brain that is responsible for storing memory and events, known as amygdala, medial temporal lobe (MTL), and parietal cortices. AD progresses in a somewhat predictable way throughout the regions of the brain [1, 2]. Longitudinal studies were done to identify and establish, independently or interactively, accumulation or deposition of amyloidal and neurodegeneration [1]. Cortical thinning in the brain areas listed above was associated with high clinical dementia rating (CDR), a test used by clinicians to access, characterize, and rate functional and cognitive performance of patients. Clinicians and researchers categorize patients with AD into one of the following:

1. Normal/healthy control (NC/HC): No impairment or decline with regard to brain functions is observed, but can in certain cases be identified through advanced imaging modalities. The positron emission tomography (PET) technique, which

measures biochemical and functional dynamics within the brain, quantifies deposits of abnormal proteins in the presence of radiotracers like Pittsburgh compound B (PiB) [3]. Such studies are very useful in study of diseases, clinical trials, and developments of new medicines, treatment and management of the disease [4]. However, PET imaging is costly and implies exposition to the radiation, which limits its use in clinical diagnosis [3].

2. Mild cognitive impairment (MCI): People are with an observable decline in performing various day-to-day activities and task. However, this decline is not sufficient enough to adversely affect one's work and relationship as a whole. Some researchers classify this MCI category further into early mild cognitive impairment (eMCI) and late mild cognitive impairment (lMCI) [5].
3. Alzheimer disease (AD): Sever decline in executing mental, physical, and social functions are observed.

The human brain and nervous system as a whole can be categorized largely in three components:

1. White matter: Tissue consisting of nerve fibers (axons), also known as tracts, surrounded by a type of sheath or covering called myelin (a kind of fatty material), which provides it a white color. It is responsible for transmitting information between gray matter regions in the brain.
2. Gray matter: Tissue consisting mostly of the brains' neuronal cell body, where neural activity takes place, present in the deep parts of the brain and in the periphery of brain. No myelin covering is found here.
3. Cerebrospinal fluid (CSF): A clear, colorless, plasma-like fluid, spread around the brain and spinal cord, which protects the brain from rigorous movement and external shock, circulates nutrients, and helps removal of metabolic waste from the brain.

Researchers have extensively studied which part of the brain is affected or involved in AD through different invasive techniques like autopsy and postmortem, and by using in vivo and noninvasive imaging modalities of magnetic resonance imaging (MRI). Time course and regions affected of the brain, as diseases progresses, vary substantially from patient to patient. However, the degenerative sequence identified through different MRI imaging techniques (i.e., MRI and SPCET) to track the progress of AD is well-documented and much agreed upon among clinicians. The same is observed even in postmortem and autopsy "Braak Maps" [6]. Neurofibrillary pathology usually starts to exhibit in the transentorhinal cortex and rapidly spreads to the entorhinal cortex before it involves the hippocampus, then spreads cortically to engulf the rest of the temporal, frontal, and parietal lobes of the brain. Furthermore, this rising deficit was asymmetric—the left hemisphere was involved to a greater extent than the right hemisphere.

A child develops their cognitive and functional abilities as it grows up, in a pattern where simple and the most basic actions progress to more complex actions. However, the reverse is observed in AD, first one starts losing the ability to perform the most complex and gradually more basic task as the disease progresses [7]. Over the past few

years, gray matter of the brain was primarily observed and considered for the study of structural atrophic changes and pathophysiological changes in AD patients. Such atrophic changes were asymmetric; in the left hemisphere, shrinkage was higher than the right hemisphere. The longitudinal study further tracked and observed patterns of cortical atrophy, gray matter loss, and built a dynamic map, which helped to distinguish among different phases of AD and differentiated AD from normal aging [6].

However, a recent neuroimaging study revealed the correlation of white matter degeneration and demyelization as an important pathophysiological feature as AD progresses [4]. For most of the neurodegenerative disorders (NDDs), including AD, the classification is based on molecular pathology—which type of protein was involved, its underlying proteinopathy, and the role of the protein processing system [8]. The study also presented an overview depicting a list of proteinopathy, proteins, diseases associated with it and presented an algorithm taking into account genes for the classification of sporadic NDDs.

2 Functional MRI

Functional magnetic resonance imaging (fMRI) places an investigational window on the brain region, to observe and investigate the working human brain in vivo of patients suffering from neurologic and psychiatric disorders. It is a methodology to detect dynamic patterns of activity in different scenarios: (1) A resting state functional magnetic resonance imaging (RS-fMRI) when the subject is not performing any intentional or instructed cognitive, language, or motor tasks, which is focused on spontaneous, low-frequency fluctuations in the blood oxygen level-dependent (BOLD) signal (2) while performing, executing, or following a scheduled task or activity, or in the presence of some external stimuli, this is detected, observed, or recorded [9].

Functional MRI offers enhanced spatial and temporal resolution compared to PET methods, and it eliminates the potential risk of ionizing radiation, so this emerged as a better method for functional neuroimaging research and study. However, PET studies thoroughly exhibit the localized nature of blood flow and metabolism changes because uses of radioactive tracer are used for this imaging technique, suggesting a relatively tight linkage among the underlying neural activity, energy metabolism, and blood flow. Therefore, it is still a preferred choice among clinicians, as it helps in diagnosing precise locations [10].

Neurons do not have any interior reserves of energy to perform their functions in the form of glucose or oxygen. When these neurons are activated or fired, they are provided with their increased demand of energy by the adjacent capillaries (blood vessels) through a process called the hemodynamic response, which in turn supplies them with an increase in regional cerebral blood flow followed by an increasing

oxygen supply. This results into change in terms of the relative levels of oxyhemoglobin and deoxyhemoglobin. This is detected by MR imaging by their differential magnetic susceptibilities, known as blood oxygen level-dependent (BOLD) contrast imaging [11]. There are multiple approaches, each with its applicability, advantage, limitations in patients with neurosurgical, neurological, and psychiatric disorders, to determine the severity and stage, to plan treatment, to study the response to medicine, and to track progress [11, 12].

The functional segregation method focuses on the local functions performed by specific brain regions and is widely used, primarily for brain mapping. The functional integration method, on the other hand, focuses on identifying the functional relationships or connectivity among different brain areas and aims to assess the brain as an integrated network (Table 1).

3 Review of Major fMRI Software Projects

Some of the widely used freely available (non-commercial) fMRI software packages and libraries are reviewed with respect to multiple parameters like operating system supported, availability of documentations, tutorials, user manuals, support, user fora, data preprocessing, noise handling, support for image processing concepts like realignment, transformation, aliasing, and registration, as listed in Table 2. Each software package listed here has its own merits and demerits, and strength and weakness, so choice largely depends upon type, quality and format of input data, preprocessing of data to handle noise, outliers, type of analysis to be done (final objective of study), number and type of features to be considered in the study, image registration, and whether other related tasks are required or not. So it is not possible to recommend one package over the other; it largely depends on the task you need to perform.

4 State of the Art: Uses of fMRI for Classification and Prediction

Identifying normal/healthy control (NC/HC) and mild cognitive impairment (MCI) (which would probably progress to AD in the due course of time), early among subjects, is extremely important to improve patients' health and to test the effectiveness of drugs by various stakeholders of the healthcare sector. One hundred and thirty-four subjects from the Alzheimer's disease neuroimaging initiative (ADNI) dataset that have both sMRI and fMRI modalities were selected. A. Abrol et al. assigned pMCI (progressive) to those who advanced to AD in the longitudinal study of three years and those who do not advance to AD were assigned sMCI (steady). The multimodal data fusion approach based on deep learning outperformed the classical unimodal machine learning-based approach; diagnostic classification and prediction

Table 1 Summary of data analysis methods applied on fMRI data for knowledge derivation

Name of method	Approach/working	Measures/models what	Advantages	Limitations
Amplitude of low-frequency fluctuations (ALFF)	It measures the total power of the BOLD signal within the low-frequency range of 0.01–0.1 Hz 0.010–0.027 Hz cortical neuronal activity, 0.027 to 0.073 Hz basal ganglia activity, 0.073–0.198 Hz physiologic noise and white matter signal	Regional brain activity	Simple Not hypothesis driven	Functional segregate
Regional homogeneity analysis (ReHo)	Measurement of the similarity among time series of a voxel under consideration and its nearest neighbors Higher ReHo → higher coherence and centrality	Synchrony of adjacent regions	Simple No a priori ROI is needed	Functional segregate
Functional connectivity density analysis (FCD)	To calculate the correlation of the BOLD time series between each voxel and all the other voxels in the brain	Aims to identify the extremely connected functional hubs of brain	No any model/assumptions are required	Does not indicate which regions are connected

(continued)

Table 1 (continued)

Name of method	Approach/working	Measures/models what	Advantages	Limitations
Seed-based (ROI) functional connectivity analysis	To compute cross-correlation among time series of the seed and rest of the brain	To find/extract regions correlated (functionally connected) with the given activity in a seed region	Simple Interpretation is intuitive	A priori determination of the seeds is required
Independent component analysis (ICA)	Use of multivariate decomposition, with the aim to separate the BOLD signal into numerous independent functional networks in the form of spatial maps, which are temporally correlated	To extract all detectable networks within the subject	Data driven, i.e., opposite to seed-based approach No any priori assumption	Presents brain networks one by one Complicated interpretation
Graph analysis	To analyze the topology of brain networks, this examines the local and global organization of neural networks To typify the entire brain map as well as different brain networks	To organize brain network optimized toward an elevated level of local and global efficiency	No any a priori assumption and minimal bias	Generally results are not insightful, and results are difficult to infer
Chronnectome	Temporal development of spatial properties of functional connectivity during the scanning	To study the spatiotemporal dynamic kind of features within brain networks	Explores dynamic functional connectivity (dFC)	Complex mathematical model and interpretation of result

Table 2 Summary of fMRI data processing software

Feature	AFNI	SPM	3D Slicer	FSL	CONN	Brain Connectivity Toolbox	fMRIPrep A fMRI Pipeline
Computing environment/OS supported	Unix	Linux and Windows	Linux, MacOSX, Windows	Mac Os X, Linux	Linux, MacOSX, Windows	Linux, MacOSX, Windows	Docker environment
Source code/programming language	C	MATLAB	C++, Python, Qt	C++	MATLAB	MATLAB	Python-nipype
Capability for preprocessing of data	–	Very Good	Very Good	Very Good	Very Good	Good	Good
Graphical user interface	Yes	Yes	Yes	Yes	Yes	Yes	Yes
First version available	1994	2000	–	2000	2010–11	2014	2016
Current/latest version	20	12	4.10.2	6.0	2018	2017	2019
Documentation, installation guide, tutorial, user guide	Good	Excellent	Excellent	Excellent	Excellent	Good	Very Good
Support for other MRI imaging modalities (structural, DTI, FDG-PET, etc.)	Elementary	Very Good	Excellent	Excellent	Good	Elementary	Structural and fMRI

accuracy are better in multimodal approach. Deep residual learning derived from the nonlinear features of structural MRI (sMRI), and extraction of fMRI features based on dynamic FC (dFC)was used to individually predict who would progress to AD from MCI [12].

Availability of different neuroimaging software and data processing tools (Table 2) may result into different results for the same set of subjects, due to differences in strategy; hence, research results may not be reproducible. Therefore, research has to integrate different modules of various software packages to fully utilize and adopt all the recent advancements made in this field [13]. The researchers here developed the fusion of neuroimaging processing (FuNP) pipeline, which is built using desired components of existing software (Workbench, FSL, FreeSurfer,AFNI),where preprocessing steps for structural (T1) weighted and fMRI is made available under the same software, such wrapper software is a good contender for novice researchers [13].

The default mode network (DMN), a very large-scale brain network formed from interacting brain regions, is known to have a highly correlated activity. When the brain is at wakeful rest, it is assumed to be a task negative network. Disruption in DMN is one of the parameters in accessing a normal versus affected person. The authors applied fusion analysis of the resting state network (RSN) for classification of a subject among NC, EMCI, and AD. Group-independent component analysis (ICA) is applied to extract the DMN and dorsal attention network (DAN) for each subject group to build a group template. Individual ICA is then applied to obtain DMN and DAN for each subject. The results achieved using this multimodal approach yield better classification accuracy and specific brain regions which get affected in AD were identified in this study [14].

The brain functional network constructed from entire time series data may overlook complex, dynamic, and interactive relationships that exist among brain regions. Zhuqing Jiao et al. proposed a multi-scale feature integration framework, which extracts time-varying, global static, and more polished features from (1) static (2) dynamic and (3) higher-order functional networks, to improve the accuracy of the classifier. The higher accuracy of the proposed multi-scale method over other single mode networks demonstrates an improvement. However, it requires a substantially large dataset to avoid overfitting [15].

Machine learning methods like support vector machine (SVM) and independent component analysis (ICA) approaches are highly guided. This requires identifying the most discriminating features from the data, which will be passed on to the classifier for classification. Recent deep learning approaches like convolutional neural networks (CNN) have a potential to identify features uniquely, which discriminate classes with very good accuracy [16]. Yan Wang et al. proposed a multimodal MR approach, where functional connectivity matrices and fractional anisotropy (FA) were used as a biomarker. Images formed through the functional connectivity network (FCN) and diffusion tensor imaging (DTI) scans of size $2 \times 90 \times 90$ were passed to a CNN deep learning model, which is composed of two convolution groups and two fully connected layers. After two convolution groups, the input image is turned into $64 \times 23 \times 23$, with 33,856 dimensions used to spread out each block into a line, so that each voxel is used as a feature. The accuracy reported using this multimodal

approach reached 92.06%, which shows an increase of 5–10% as compared to the unimodal approach [16].

There is a subtle difference between the MCI and the AD group, which makes the task of classification a challenging one. At the same time, it is very important to differentiate people who surely will progress to AD in the due course of time from the MCI stage and those who will not progress to AD. Saman Sarafet al. presented a deep learning-based algorithm built upon an optimized convolutional neural network (CNN), which classifies NC, MCI, and AD patients using structural and fMRI data (multimodal approach). The minimum accuracy of three class classifier was reported to be 99% using sMRI and 96% in fMRI pipeline [17].

There is inherent unevenness in the brain connectivity for each individual, so the connectivity between two brain regions alone may not be adequate to distinguish NC and MCI or AD subjects. Based on this premise, brain network connectivity pattern analysis, which aims to look for faint changes within the pattern of connectivity among multiple or all regions in the sub-network, may increase classification accuracy. Based on this premise by Z. Wang et al. regions of interest (ROIs), a sub-network within the DMN was selected, then calculated the correlation coefficients between all the possible ROI pairs, and formed a feature vector for each subject. Next they proposed a regularized linear discriminate analysis (LDA) approach with the aim to reduce the noise effect it has on the accuracy of the classifier. After this a decision tree-based multi-class AdaBoost classifier is applied. The results achieved an accuracy of 75.8% with regularized LDA and AdaBoost. The results demonstrate that the classification algorithm may fail to distinguish MCI patients as AD or NC subjects, because in MCI patients some of the observed functional characteristics and connectivity patterns are shared, as they are similar to that found in NC and AD patients [18].

Neuropsychological assessment tests done by the clinicians were the first line of action in diagnosis and prediction of AD until various noninvasive MR imaging modalities were developed and used to their fullest potential. Recently, computer-based decision support systems augmented with the application of machine learning, artificial intelligence, and deep learning were introduced, which are more objective and sensitive to detecting subjects in the early stage of AD. Soroosh Golbabaei et al. used rs-fMRI data to derive FCN, and various graph-based biomarkers were used to build FCN. Correlation among the brain network (regions of brain), values observed among this graph-based biomarkers (1) eigenvector centrality (2) betweenness centrality (3) local efficiency (4) node strength (5) local clustering coefficient (6) local characteristic path length, and (7) five different AD physiological assessment test scores are evaluated, which exhibited a positive correlation with mini mental state examination (MMSE), clinical dementia rating (CDR), and the neuropsychiatric inventory questionnaire (NPI-Q) assessment test [19].

Machine learning-based methods like support vector machine (SVM) have been explored extensively in many studies by various researchers to classify AD, MCI, or HC patients. The accuracy of such single SVM is less than 90%. To improve the accuracy, researchers have proposed to use the random support vector machine cluster, preceded by rigorous preprocessing, which at random selects the samples and

the features to construct multiple SVMs. The random SVM cluster is then formed from these multiple SVMs and is used for feature selection and classification. The classification accuracy reported by this approach was 94.44% [20].

Table 3 provides a summary and critical review of various work and studies on computer-aided detection systems, which used fMRI imaging modality to classify Alzheimer's disease (AD). The table covers and answers questions about the objective or purpose achieved, dataset used for the study, sample size and distribution used, whether it is a longitudinal study or not, among which classes the used classifier classifies (two class or multi-class classifier), whether other psychological and physiological parameters considered or not, which method, approach, technique, and biomarkers are used, results and accuracy reported, and software or frameworks used for data processing, etc., are addressed.

5 Conclusion and Future Work

From various in vivo and postmortem studies, it is found that region of brain and area affected in brain for the patients of Alzheimer disease, in its later stage (brain atrophy related changes), are found to be same across different patients. However, in its early stage, AD showed signs of involvement of different brain regions with deferment in time and a different order and magnitude for different people. There are diverse MR imaging modalities (technique) which facilitate clinicians to analyze or study, make diagnoses, monitor disease progress, and decide surgical intervention and other alternative lines of action to control the spread and/or to slow down irreversible damage, each with its own peculiarities, advantages, feasibilities, and limitations in the treatment of neurodegenerative diseases.

To mention few of them biomarkers, fractional anisotropy (FA) and mean diffusivity (MD) derived from diffusion tensor imaging (DTI) give quantitative evaluation of "*tissue integrity*". Blood oxygen level-dependent (BOLD) signal/response, used as a biomarker derived from fMRI experiments estimates "*brain activity*" among different brain region. Positron emission tomography (PET) scans in the presence of radioactive tracer Fluorodeoxyglucose (FDG) is an indicator of *glucose hypometabolism,* i.e., slow or reduced rate of metabolism in Alzheimer's patient. So in general these techniques are good contender to detect those subtle degenerative changes in the brain structure (in the earlier stages of disease) even before any notable symptoms are observed or seen in the patient. Further it is a good choice and probability that when this multimodal approach is used in combination with train the classifier, i.e., to use more than one biomarkers, derived/computed by applying more than one MR technique for computer-aided detection of Alzheimer's disease yields in better sensitivity and specificity, resulting in better overall efficiency. As different classifiers are trained using different parameters/features, they will have different decision boundaries and each makes different error. The idea is to use those different decision boundaries, which reduce errors and thus improve overall accuracy.

Here we suggest a novel approach, as shown in Fig. 1, which computes different

Table 3 Literature review of various fMRI-based studies with respect to objective, dataset, sample size, classes considered, method and biomarkers used, result/accuracy reported, type of classification algorithm used, and software toolbox or framework used

Sr. No.	Objective/purpose	Dataset used—own or third party	Sample (Dataset) Size & characteristics	Longitudinal study?	Stages considered	Other parameters included? (MMSE CDR, etc.)	Method/approach/technique/biomarkers used	Result/accuracy reported	Classifier used is two state or multi state	S/W package/tool used
[12]	To test hypothesis, multimodal data analysis improves diagnostic classification and prediction accuracy	ADNI	134	Yes—3 years span	MCI & AD	NO	Dynamic functional connectivity(dFC) and structural MRI (sMRI), group ICA, cross-validation	Improvement in accuracy, sensitivity, specificity and balance accuracy	Two states	SPM
[13]	To design →A wrapper software built upon existing soft wares →User-friendly GUI, fully automated one, volume and surface-based preprocessing pipelines →Reproducible results									
[14]	To derive DMN and DAN in RSN (rs fMRI), to detect informative biomarkers	ADNI	54 (18 for each of the three group)	NO	NC, EMCI, AD	Reported but not used as parameter in classifier	Combining ICA, group analysis, fusion analysis to classify subjects	Improvement reported	Two states	–
[15]	Proposed multi-scale feature integration framework to improve classification accuracy	ADNI Go & ADNI2	60 30—eMCI 30—NC	NO	eMCI & Normal control	NO Only f-MRI modality is considered	Global static, time-varying, more refined feature are derived	Balanced accuracy 91.13%	Two class classification	SPM

(continued)

Table 3 (continued)

Sr. No.	Objective/purpose	Dataset used—own or third party	Sample (Dataset) Size & characteristics	Longitudinal study?	Stages considered	Other parameters included? (MMSE CDR, etc.)	Method/approach/technique/biomarkers used	Result/accuracy reported	Classifier used is two state or multi state	S/W package/tool used
[16]	To use multimodal data and apply deep learning approach to improve accuracy	Beijing Xuanwu Hospital	105 35—AD 30—aMCI 40—NC	NO	AD, aMCI, NC	Yes DTI/DWI & fMRI	Deep learning—CNN, functional connectivity and fractional anisotropy	Balanced accuracy 92.06%	Three class classification	DPASR, PANDA, GRETNA, toolbox
[17]	To design an deep learning-based algorithm/topology, built upon optimized CNN, which predicts MCI to AD conversion	ADNI	275-fMRI (52-AD, 92-NC, 131-MCI) 1076-sMRI (211-AD, 91-NC, 774-MCI)		NC, MCI, AD	sMRI and f-MRI Multimodal approach	Applied CNN to improve classification accuracy sMRI and fMRI pipelines were developed	sMRI → 99.77% ± 0.36% fMRI → 97.5% ± 1.16%	Three class classifier	FSL
[18]	To design a classifier using rs-fMRI data based on FC and DMN, handles noise	Subjects were recruited (own data), approved by MST	10—AD 11—MCI 12—HC	NO	NC, MCI, AD	NO	Brain network connectivity pattern analysis in ROI, LDA to handle noise, AdaBoost	75.8% With regularized LDA, performs well on limited data size	Three class classifier	Freesurfer AFNI
[19]	To find the AD biomarkers recognizable through the graph based measurement and to derive correlation	ADNI Of it with AD physiological test score	21—AD 21—NC	NO	NC, AD	NO	FCN was constructed Graph-based biomarkers were derived and evaluated	Graph-based biomarkers value in specific brain regions were compared	NA	FSL
[20]	To build multiple SVM from randomly selected samples and features from each cluster	ADNI	25—AD 36—NC	NO	NC, AD	NO	Random selection of samples and features, repeated k times to form k SVM, then random SVM cluster is created	Accuracy → 94.44%	Two class classifier	DPARSF—data processing assistant for rs-fMRI

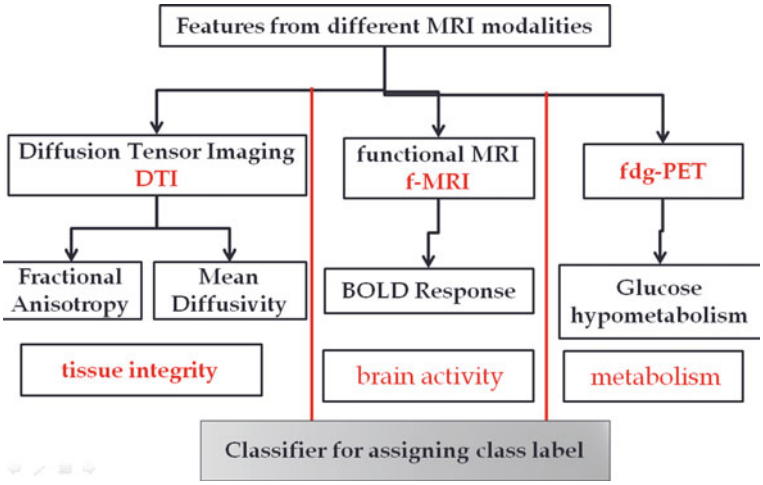


Fig. 1 Proposed MR multimodal approach for computer-aided detection system for Alzheimer's disease

quantitative biomarkers, among the patients, healthy controls. These quantified values derived from these biomarkers are then used as a parameter or feature to train the classifier for the three different MR modalities and techniques discussed earlier.

References

1. d'Oleire Uquillas F et al (2018) Interactive versus additive relationships between regional cortical thinning and amyloid burden in predicting clinical decline in mild AD and MCI individuals. NeuroImage Clin 17(September 2017):388–396. https://doi.org/10.1016/j.nicl.2017.10.034
2. Ledig C, Schuh A, Guerrero R, Heckemann RA, Rueckert D (2018) Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. Sci Rep 8(1):1–16. https://doi.org/10.1038/s41598-018-29295-9
3. Rahim M, Thirion B, Comtat C, Varoquaux G (2016) Transmodal learning of functional networks for Alzheimer's disease prediction. IEEE J Sel Top Signal Process 10(7):1204–1213. https://doi.org/10.1109/JSTSP.2016.2600400
4. Nasrabady SE, Rizvi B, Goldman JE, Brickman AM (2018) White matter changes in Alzheimer's disease: a focus on myelin and oligodendrocytes. Acta Neuropathol Commun 6(1):22. https://doi.org/10.1186/s40478-018-0515-3
5. Cui X et al (2018) Classification of alzheimer's disease, mild cognitive impairment, and normal controls with subnetwork selection and graph kernel principal component analysis based on minimum spanning tree brain functional network. Front Comput Neurosci 12(May):1–12. https://doi.org/10.3389/fncom.2018.00031
6. Thompson PM et al (2003) Dynamics of gray matter loss in Alzheimer's disease. J Neurosci 23(3):994–1005. doi:23/3/994 [pii]
7. Manuscript A (2011) Tracking Alzheimer's disease. https://doi.org/10.1196/annals.1379.017.tracking

8. Kovacs GG (2014) Current concepts of neurodegenerative diseases. Cit EMJ Neurol 1:78–86
9. Megan JSS, Lee H, Smyser CD (2013) Resting state fMRI: a review of methods and clinical applications. Am J Neuroradiol 9(5):6–9. https://doi.org/10.3174/ajnr.a3263.resting
10. Buxton RB (2015) The physics of functional magnetic resonance imaging (fMRI) Richard. IOP Publ Ltd. 76(9):1–20. https://doi.org/10.1088/0034-4885/76/9/096601
11. Lv H et al (2018) Resting-state functional MRI: everything that nonexperts have always wanted to know. Am J Neuroradiol 39(8):1390–1399. https://doi.org/10.3174/ajnr.A5527
12. Abrol A, Fu Z, Du Y, Calhoun VD (2019) Multimodal data fusion of deep learning and dynamic functional connectivity features to predict Alzheimer's disease progression. 2019 41st Annual international conference of the IEEE engineering in medicine and biology society, pp 4409–4413. https://doi.org/10.1109/embc.2019.8856500
13. Park BY, Byeon K, Park H (2019) FuNP (fusion of neuroimaging preprocessing) pipelines: a fully automated preprocessing software for functional magnetic resonance imaging. Front Neuroinform 13(February):1–14. https://doi.org/10.3389/fninf.2019.00005
14. Pei S, Guan J, Zhou S (2019) Fusion analysis of resting-state networks and its application to Alzheimer's disease. Tsinghua Sci Technol 24(4):456–467. https://doi.org/10.26599/TST.2018.9010099
15. Jiao Z, Xia Z, Ming X, Cheng C, Wang SH (2019) Multi-scale feature combination of brain functional network for eMCI classification. IEEE Access 7:74263–74273. https://doi.org/10.1109/ACCESS.2019.2920978
16. Wang Y et al (2018) A novel multimodal MRI analysis for Alzheimer's disease based on convolutional neural network. In: Proceedings annual international conference of the IEEE engineering in medicine and biology society EMBS, vol 2018-July, pp 754–757, 2018. https://doi.org/10.1109/embc.2018.8512372
17. Sarraf S, Desouza DD, Anderson JAE, Saverino C (2019) MCADNNet: recognizing stages of cognitive impairment through efficient convolutional fMRI and MRI neural network topology models. IEEE Access 7(Mci):155584–155600. https://doi.org/10.1109/access.2019.2949577
18. Wang Z, Zheng Y, Zhu DC, Bozoki AC, Li T (2018) Classification of Alzheimer's disease, mild cognitive impairment and normal control subjects using resting-state fMRI based network connectivity analysis. IEEE J Transl Eng Heal Med 6(July):1–9. https://doi.org/10.1109/JTEHM.2018.2874887
19. Golbabaei S, Dadashi A, Soltanian-Zadeh H (2016) Measures of the brain functional network that correlate with Alzheimer's neuropsychological test scores: An fMRI and graph analysis study. In: Proceedings annual international conference of the IEEE engineering in medicine and biology society EMBS, vol 2016-Oct, pp 5554–5557. https://doi.org/10.1109/embc.2016.7591985
20. Bi XA, Shu Q, Sun Q, Xu Q (2018) Random support vector machine cluster analysis of resting-state fMRI in Alzheimer's disease. PLoS ONE 13(3):1–12. https://doi.org/10.1371/journal.pone.0194479

A Pilot Study of Indian Ragas-Based Music Therapy for Enhancement of Chronic Patient Health Condition

Milind U. Nemade

Abstract International Diabetes Federation said that there is 5% of Indian population agonizes from diabetes, and it will rise to 123 million by 2040, also one study reveals that three Indians in ten suffered from hypertension and 89% of Indian population suffered from stress. There are various parameters which will impact on the medical health of a patient, and music is one of them. Emotions of individual by changing the timbre will affect by raga as said by ancient Indian classical musical experts some decades ago. Due to advancement in technology and day-by-day increasing work pressure on a human being due to competitions, there is need of an hour to improve health and reduces stress by other traditional way. This review work explores raga music therapy as scientifically the traditional way of medicine. A pilot study reveals that the raga music therapy will enhance health of diabetics, low BP, high BP, and stress of chronic disease patients. At the end mobile app development process flow is explained on the similar concept which may greatly help in improvement of different chronic patients' health.

Keywords Emotions · Raga and rasa · Music therapy · Hypertension · Diabetes

1 Introduction

In our standard of living, person generally stumbles upon stress leading to various physical and mental illnesses. International Diabetes Federation said that there is 5% of Indian population agonizes from diabetes, and it will rise to 123 million by 2040, also one study reveals that three Indians in ten suffered from hypertension and 89% of Indian population suffered from stress. One experiment conducted on hypertension patient shows listening of low arousal music for one month will greatly reduce heart rate and mean arterial pressure compared to impact of drug [1].

Old Indian tradition provides various healings like yoga, meditation, and raga healing. Raga healing was an old traditional method which helps in enhancement

M. U. Nemade (✉)
K.J. Somaiya Institute of Engineering and Information Technology, Ayurvihar Complex, Everard Nagar, E.E. Highway, Sion-E, Mumbai 400022, India
e-mail: mnemade@somaiya.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_27

of health by raga healing method. Raga darbari, pooriya, khamaj, and kanada help to solve mental stress problem of fits. Raga Malhar also known as Miya Malhar useful on asthma and sunstroke medical treatments soothes anger and control mental instability, raga jaijaivanti helpful in comforting mind and curing mental disorders. Raga as rasa and it means color or mood with specific musical instructions in Sanskrit [2]. Music therapy is totally depends on the right notations and properly use of nada, shruti, swara, raga, tala, and laya elements [3]. There is a time of the day and season to perform Indian music classical raga. It suggests specific moods within the listener. Each raga in music therapy produces rasa which consist of one or more emotions out of seven emotions. Usually, music conveys seven emotions as strength or courage, romance, peace, devotion, sadness, dispassion, and anger. Each raga produces a singular emotion (rasa) consisting of one or more of those emotions [4]. Here, machine learning may also be a very important part which science and theories have defined to be useful notes for curing diseases. Hence, it may be useful for the therapy sessions [5].

1.1 Healing Diseases with the Music Therapy

Raga has a collection of different unique notes that are inclusive of special properties. These ragas are divided into two systems North Indian music and South Indian music. Seven swaras Sa, Re, Ga, Ma, Pa, Dha, and Ni are very useful and used for the saptak or scale [6]. Flat and sharp swaras used for lowering and increasing pitch of the music. There are 12 notes or swaras present in each of the ragas [7]. One study reveals that Sarang family ragas, Kalyani, and Charukesi have capability to cure heart vein blockages and heart diseases. Raga Bageshri and Jaijayanti work well for type 1 and type 2 diabetes patients. During oncology chemotherapy sessions, playing raga Bhairavi and Lalit will give good results on patient health. Medical patients recover fast compared to patient taking chemical drug by using raga music therapy [8]. EMOTIV wireless EEG brain wear hardware and machine learning algorithms software can be used for stress analysis. Raga with above hardware and software will track the sleep of a person. One research paper said that frustration and stress cause hypertension and less concentration problem of patients in a day time if the human body gets sufficient sleep that is required. Listening raga will overcome this problem, and patient will get proper sleep and can avoid disease like hypertension [9].

1.2 Ragas and Emotions

Pitch collection and characteristic phrase are two properties of raga, and they are related to moods with its pitches. Swara “Pa” synchronizes our body and also our mind too. Swara helps the excretory organs and reproductive systems of our body.

Table 1 Raga details for disease treated

Sr. No.	Name of raga	Inventor name	Time of listening raga	Disease treated
1	Bageshri	Miyan Tansen	9 PM to 12 AM	Diabetes
2	Chandrakauns	Godess Parvati	9 AM to 12 PM	Heart ailments and diabetes
3	Puriya Dhanashree	Bhat Khandeji	4 PM to 7 PM	Acidity and high blood glucose
4	Darbari	Miyan Tansen	10 PM to 1 PM	Easing tension
5	Hindol	Bhat Khandeji	3 AM to 7 AM	hypertension, blood purification, rheumatic arthritis and backache
6	Todi	Bhat Khandeji	8 AM to 10 AM	Relief from high blood pressure
7	Ahirbhairav	Lord Shiva	6 AM to 9 AM	Indigestion, rheumatic arthritis, and hypertension
8	Bhupali	Bhat Khandeji	7 PM to 10 PM	High blood pressure, headache and relief from cold
9	Malkauns	Goddess Parvati	12 AM to 3 AM	Low blood pressure
10	Asawari	Bhat Khandeji	7 AM to 10 AM	Low blood pressure and builds confidence

In mind, Apana feels the emotions and gains a control over or get rid of them. Both Prana and Apana work in opposite directions. They reinstate symmetry in our body and mind [10]. Selected quite emotions conveyed by pitches occur in ragas. The flavor or mood of the raga conveys by the raga pitch and its characteristic phrase. There are 11 basic moods in North Indian mix of music, dance, and poetry which will be portrayed through a karuna, shringar, shanta, veer, raudra, vyragya, bhakti, bhayanak, hasya, bachata, and adbhuta. Ragas will also help in Carnatic music. Raga Bhairav helps effectively in the diseases for the skin. Raga Charukesi is used to rouse someone from deep sleep.

Table 1 shows name of raga, inventor name, time of listening raga, and disease treated details of different ragas used to cure different diseases which can be heard at different time of the day.

1.3 Raga and Time

Raga melodies perform certain times of the day and night. It is unique characteristic of Indian classical music [11]. Raga Megha and raga Malhar are related to the time

of year, raga Basant related to the autumn season and raga Bahar related to the spring season. During season, patient can sing and play seasonal ragas at any time of the day and night. This connecting time of the day or night, with the raga, depends on daily cycle of changes that occur in our own moods and emotions which regularly changes during this different moment of the day and stimulates different moods and emotions [4]. The target of this study is to research the effect of Indian genre on brain activity during normal relaxing conditions using electroencephalography (EEG) [12]. Given a music set that can be considered, we initially have to extract the exact predominant pitch values at every instinct point of the time to be referred, also we have to convert these to cents scale and then map them to a single octave. After these processes, we can then identify the stable note regions to construct pitch class profile and the n-gram distribution [13, 14]. These pitch class and the n-gram distribution are very different and need to be notified on very high scale basis.

2 Related Works

One research paper said a common clinical disease is hypertension, and it is a major risk to human health. In the world, an estimated 26% of the world's population has hypertension and the likely to increase 29% by 2025. The high occurrence of hypertension demands a tremendous public health burden. Around 9.3% of the global adult population suffered from diabetes in 2019 and rise to almost 11% by the year 2045 [15]. Diabetes mentions to a group of metabolic disorders that result in chronic high blood sugar levels, and it can lead to serious health problems, such as cardiovascular disease, chronic kidney disease and stroke. In the world, 40% adults aged 25 and over have high blood pressure in 2018 [16]. It is observed in many clinical findings that specific types of music can reduce blood pressure (BP). For an anti-hypertension treatment, music therapy played a vital role. In some research paper, authors proposed method to detect the current BP value with a measurement technique. According to the current BP value, to ease the hypertension of patients, the MP3 player selects specific types of music for playing in order [4, 17, 18]. One paper explained an electroencephalogram (EEG) test that detects electrical activity in your brain using small, metal disks attached to the scalp. Brain cells communicate via electrical impulses and are active all the time, even when you are asleep. This activity shows up as curly lines on an EEG recording. But the stress is the common thing which can be recognized with brainwave using EEG technique, simply by placing the electrodes on scalp. In one paper, authors implemented a system which will generate music according to user needs and specifications. This system uses genetic algorithms approach and Java language for generating music by inputting swaras of ragas [9]. Table 2 shows the summary of previous works on raga music therapy.

Table 2 Summary of previous works on raga music therapy

Ragas	Diseases treated	Measurement tools/parameters	Intervention type	Period/frequency	Observations	Conference/journal details	Outcomes
Raga Chandrakauns	Diabetes	Questionnaire to patients, BSL measured for pre-exposed and post to music exposed	Two group of persons formed evaluated 50 diabetes patients (60–75 years) one group informed to listen music of their choice and other group not exposed to music	Half an hour for 3 months	Patient exposed to music found relax and happy than other group	9th Diabetologist Conference Dallas, Texas, 2016	Reduced BSL
Raga Bhagershri/ Raga Jaijayanti	Diabetes	Estimated random blood sugar before and after the test	59, type 2 diabetes persons of different age group and different duration of diabetes exposed to listening of raga music	30 min	Patients feel relaxed for the period the music was heard	Music and Medicine Interdisciplinary Journal, Israel, 2018 [19]	Significant reduction in blood sugar
Raga Todi/Ahirbhairav	Hypertension	MP3 player detect BP value with cuffless measurement method	15 old persons having systolic blood pressure higher than 140 mmHg exposed to listen selected music	25 min per day for 4 weeks	Measure users' BP value, and this value determines the type of music	31st Annual International Conference of IEEE, Minneapolis USA, 2009 [15]	Improve hypertension

(continued)

Table 2 (continued)

Ragas	Diseases treated	Measurement tools/parameters	Intervention type	Period/frequency	Observations	Conference/journal details	Outcomes
Raga Malkauns	Hypertension	HR and RR and MAP parameters	200 elderly hypertensive patients exposed to Indian classical music	Short period everyday for one month	Lower difference in HR and MAP was recorded after listening music. Mood assessment performed using PANAS scale	Research article Current Science 2020 [1]	Low HR and MAP value recorded and positive mood changes of patients found
Raga Malkauns	Hypertension	Spectral flux and spectral centroid parameters	Two groups study and control formed and exposed to 25 samples per group	Twice a week 30 times, 15 min over a period of 3 months	Reduction in value of SBP after exposed to music therapy treatment	ICCUBEA, 2017	Raga Malkauns recorded different effects on high and low BP measurements
Raga Darbari	Stress	EEG	Three electrodes placed in one of the brain lobe and EEG output recorded	30 min	Beta brain wave of stressed patients changes to alpha brain wave	ICCPEIC, 2016 [20]	Reduction in stress and patient feels better

(continued)

Table 2 (continued)

Ragas	Diseases treated	Measurement tools/parameters	Intervention type	Period/frequency	Observations	Conference/journal details	Outcomes
Raga Darbari/Raga Bhimpalasi	Stress	Kernel density estimation and MLP for emotion detection	Eight electrodes placed with reference to earlobes, EEG output recorded after listening raga music	30 min	Visualize stress perception and EEG feature extracted using KDE and MLP	ICACEA, 2015 [21]	Reduces stress and obtained accuracy of emotions using NN

3 Indian Raga-Based Music Therapy Approach

Based on former research, Indian raga-based music therapy approach includes listening of raga music, measurement and monitoring of blood pressure, and blood sugar level and stress monitoring using EEG signal analysis are a need of the hour. This approach required five different modules; listening raga music, blood pressure measurement, diabetes BSL measurement, EEG measurement and analysis of recorded data, and monitoring enhancement of chronic patient health condition. Block diagram of Indian raga-based music therapy for improvement of chronic patient health is shown in Fig. 1. Before applying this therapy to the patient, he should choose quiet and pleasant place for listening raga music. He should sit with comfort for 5 min. Then the patient should record BSL if he is a diabetic patient or record systolic and diastolic value if he is a BP patient or record wave frequency value in Hz if he is a stress patient. Then he can select raga music for listening based on the category of patient he was chosen. Note that there is time of listening every raga music as mentioned in earlier section of this paper when that raga music will be effective. To get maximum health improvement, listen respective raga in the specified time in Table 1. After the raga music is played, the method again records reading and compares both the readings before and after the treatment is been done.

3.1 Blood Pressure Measurement Module

Blood pressure is recorded as two numbers: The systolic pressure as the heart beats over the diastolic pressure as the heart relaxes between beats. The measurement is written one above or before the other, with the systolic number on top and the diastolic number on the bottom. Blood pressure measurement of 120/80 mmHg (millimeters of mercury) is expressed as "120 over 80." Many references show that music is a

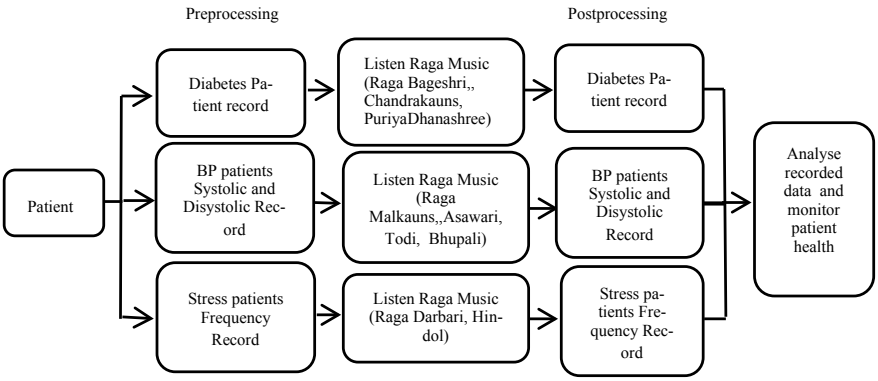


Fig. 1 Block diagram of Indian raga-based music therapy

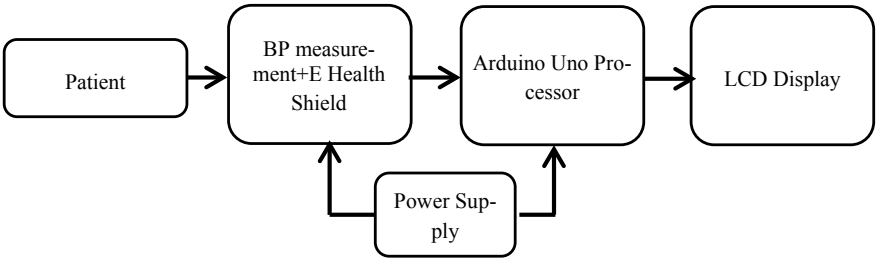


Fig. 2 Blood pressure measurement module

remedial solution in treating many body disorders. Blood pressure and heart rate get affected by music signal, and these features of raga music therapy reduce the risk of developing cardiovascular disease. In current competitive world, it is a critical task that raga music therapy is maintaining the respiratory rate within the normal range. In order to reduce risk associated with cardiovascular diseases, use of Indian and Western music has been considered among many preventive programs for long time [22]. Simple blood pressure measurement module using Arduino UNO processor is shown in Fig. 2.

3.2 Electroencephalography (EEG)

Electroencephalography (EEG) exploring neural activity and brain functioning measures electrical activity generated in the brain. The physiological signals of the both control as well as experimental groups were recorded before and after a timespan month's period. The method for measuring raw EEG signals was recorded for a period of 15 min. using three lead EEG electrodes [23]. For stress measurement, EMOTIV wireless EEG brain wear hardware with machine learning algorithm can be used. EEG signal is measured using electrodes placed on the scalp mentioned in another method. It records the electrical field generated by the nerve cells changes in the brains' electrical activity. Very high time resolution is required to determine the precise moments at which these electrical events take place [4]. Figure 3 shows patterns of EEG signal. There are four basic types: Delta, Theta, Alpha, and Beta waves. Delta wave is observed in deep sleep or deep meditation state and has 0.5–4 Hz frequency range. Theta wave is observed in drowsiness state and has 4–8 Hz frequency range. It detects using computational methods. Alpha wave is associated with state of peace and observed in relaxation state play important role to distress a person. It has 8–12 Hz frequency range [24]. Beta is fast wave observed in large active thinking, attentive people, peoples engaged in problem solving, judgment, decision making, and engaged in focused mental activity state and has 12–30 Hz frequency range [25].

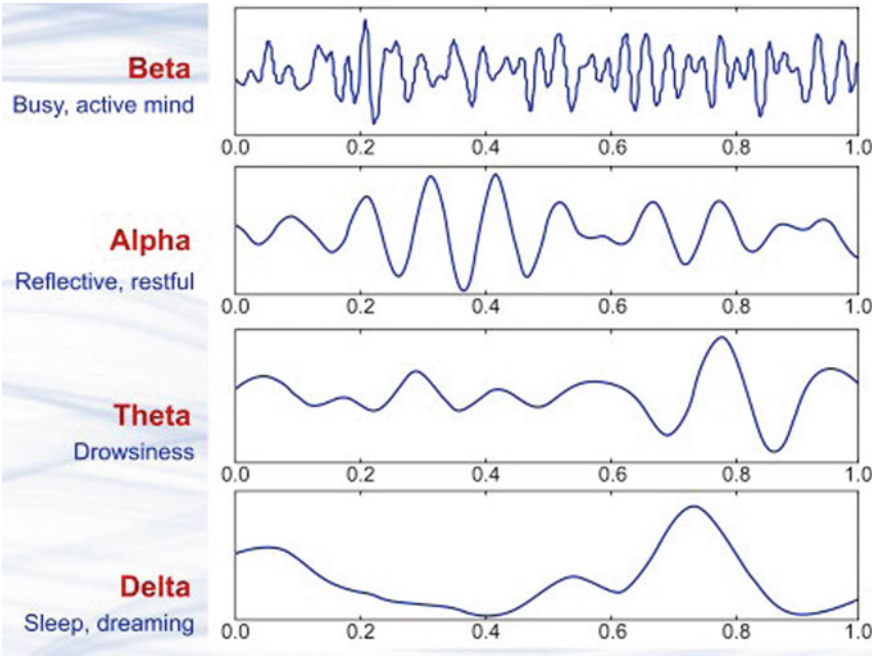


Fig. 3 EEG signal waves

3.3 Stress Reduction Using Raga Music Therapy

We know blood pressure patients under stressed condition required daily medicine to control it. The main purpose here is to detect and control the stress level of a person using raga music therapy. The stress level of the human will be detected based on the captured and analyzed brain EEG signals. In the method, digital filter will be used for noise reduction. After recording stress level status of the person, corresponding raga music shall be played to reduce the stress level of a person, which control the stress level of the subject.

Figure 4 shows the stress reduction process flow. The entire layout structure shows how the brain cells work and how raga music helps the human brain to reduce stress. Total 19 electrodes are used for measurement of stress. Placement of electrodes is like on the right side of the human skull, five electrodes are placed, first electrode at 10% location from the Nasion, then second to fourth electrode placed at every 20% location, and fifth electrode at 10% location from fourth electrode. Similarly on the left side of the human skull, five electrodes are placed in the same manner. Three electrodes are placed at center location of brain at equal distance. Three electrodes are placed at equal distance between right side of the human skull and center placement location; similarly other three remaining electrodes are placed at equal distance

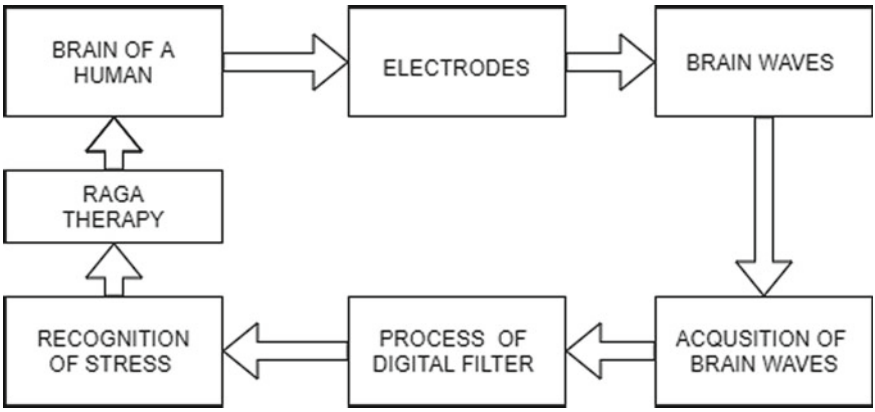


Fig. 4 Stress reduction process flow

between left side of the brain and center location. This placement of electrodes gives effective calculation of the amount of stress.

3.4 Diabetes BSL Measurement Module

Diabetes mellitus is an incurable disease resulting from an insufficiency of insulin in the body, causing high blood glucose levels, known as hyperglycemia. It is very difficult to treat diabetes mellitus disease because it is not only linked with a single organ, but with the whole composition of the body. There is no treatment found so far for this disease, but we can make attempt to bring down blood sugar level by using different therapies. Raga music therapy for diabetes mellitus releases the mind of tension and stress, normalize blood circulation, enhance metabolic function, and improve immune system of the body. All kinds of music can be used in raga music therapy for diabetes mellitus.

As shown in Fig. 5, a glucose meter determines the approximate concentration of

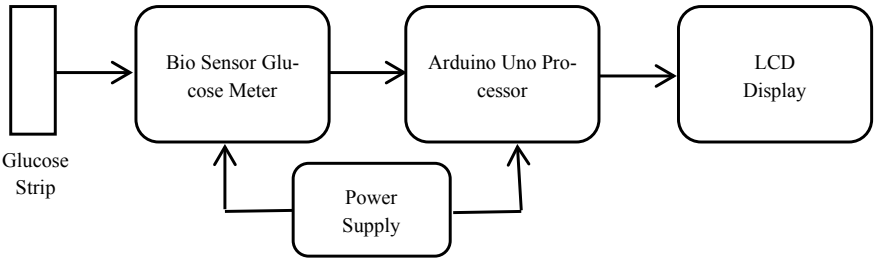


Fig. 5 Diabetic BSL measurement module

Table 3 Glucose comparison between healthy and diabetic patients

Type	Fasting (Mg/dl)	Post meal (Mg/dl)	2–3 h after meal (Mg/dl)
Normal	80–100	170–200	120–140
Pre-diabetes	101–125	190–230	140–160
Diabetes	126+	220–300	200+

glucose in the blood. It can also be a strip of glucose paper dipped into a substance and measured to the glucose chart. The hardware and software features of Arduino UNO-based system are used to determine the approximate concentration of glucose in the blood. Biosensor is used in diabetes BSL measurement module for monitoring blood glucose levels. Table 3 shows the glucose parameters for healthy patients and the diabetic patients. It shows that diabetic patients require more amount of glucose compared to the regular patients. We can also use mobile app consisting of different types of raga music. It helps to improve patient health. One such example is DeMuse application which focuses music on stress [26].

4 Mobile App Development Process Flow

As shown in Fig. 6, raga music therapy-based mobile app development flow is need of an hour today. Motivations for development of such mobile app are the review carried out for raga music therapy. The advantage of developing such mobile app is the ease of using it. Every patient will download such mobile app in their mobile and use it remotely. Patient need not to approach any music therapy center for treatment. On the basis of measured value inserted in to the app before listening and after listening of raga-based music, blood sugar level and other parameters will analyze and monitor as mentioned in earlier sections.

5 Discussion

Review work presented in this paper explained that how raga music therapy can be useful for maintaining and monitoring health of chronic patient diseases. Main focus was to identify how raga music therapy will be helpful for enhancement of health of the diabetes, blood pressure, and stressed patients. Observation is that nowadays due to technological advancements, machine learning approach proved to be effective tool to solve many problems in different fields of engineering. We can use this approach to identify which raga in raga music therapy will be more useful for patient health enhancement. Remote monitoring and enhancement of patient health with the help of mobile app and applying machine learning algorithms on the data received from different patients and analyzing of that data will be the future research work in the

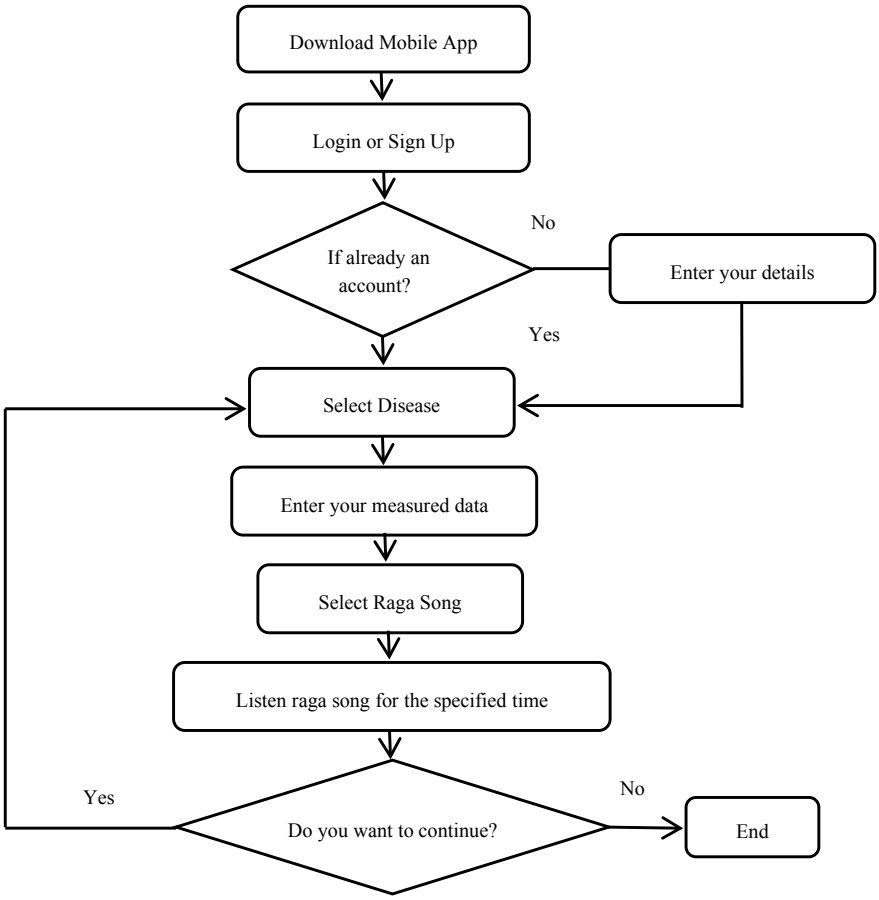


Fig. 6 Mobile app development process flow

raga music therapy. One can also work on development of a device which consists of blood sugar measurement, blood pressure measurement, and EEG measurement module along with abovementioned approach. This will greatly help chronic patients and health supervisor to monitor patient health remotely.

6 Conclusion

This review work addresses the objectives and aims to present a healthy circumstance which is able to help healthcare provider to treat patient by injecting music spur, and resulting data analysis would be helpful to determine the improvising effect in the patients' health condition. Raga music therapy method proposed in this paper will

enhance health of diabetics, low BP, high BP, and stress chronic disease patients by changing the timbre of human body. Raga music therapy is a need of the day since current advances in technology and rising of working load on human being is accompanied by stress. Patients have difficulties attending on sight raga music therapy may also participate in remote session method for raga music therapy. We explained here mobile app development process flow on the similar concept which may greatly help in improvement of different chronic patients' health.

References

1. Shankar VM, Geethanjali B, Veezhinathan M, Hariharakrishnan J, Balakrishnan N, Lakshmi L (2020) Evaluating the effect of music intervention on hypertension. Curr Sci 118(4)
2. Jennifer H (2020) The history, current role, and future of music therapy in India: international interviews with Prof. Dr. Sumathy Sundar and Aastha Luthra. Voices World Music Ther 20(1)
3. Ray LC, Yuka F (2019) Machine learning and therapeutic strategies in VR. In: ACM proceedings in Artech
4. Kurizuka Y, Miyake Y, Kobayashi Y (2004) CO-creative walking support as music therapy. In: Proceedings of the 2004 IEEE international workshop on robot and human interactive communication, Kurashiki, Okayama Japan. IEEE
5. Mounika A, Srilatha B, Neetha K, Desai KV, Radha Kishan N, Aparna B (2014) Indian music therapy: could it be helpful in the management of mental depression? Int J Med Res Health Sci (IJMRSH) 3(2):354–357
6. Katte T, Tiple BS (2014) Techniques for Indian classical raga identification—a survey. In: Annual IEEE India conference (INDICON). IEEE
7. Nagarajan K, Srinivasan TM, Nagendra Hongasandra R (2015) Immediate effect of listening to Indian raga on attention and concentration in healthy college students: a comparative study. J Health Res Rev 2(3):103–107
8. Todiruta A, Goga M, Pestritu L, Ciobanu L, Nicolae G (2013) Stress free—a tool for stress determination and treatment through computer music generation. In: IEEE International conference on healthcare informatics
9. Vasudha J, Iyshwarya G, Tamarai Selvi A, Iniyaa S, Jeyakumar G (2011) Application of computer-aided music composition in music therapy. Int J Innov Manage Technol (IJIMT) 2(1)
10. Balaji Deekshitulu PV (2018) Indian Ragas for treating health problems. Int J Music Film (IJMF)
11. Lele J, Aditya A (2019) Towards Raga identification of Hindustani classical music. In: IEEE Pune section international conference MIT World Peace University. IEEE
12. Phasukkit P, Mahrozeh N, Kumngern M, Supan T (2015) A simple laboratory test of music therapy for insufficient sleep. In: The 2015 biomedical engineering international conference (BMEiCON-2015). IEEE
13. Kumar V, Pandya H, Jawahar C (2014) Identifying Ragas in Indian music. In: 22nd International conference on pattern recognition. IEEE
14. Sharma H, Bali RS (2014) Raga identification of Hindustani music using soft computing techniques. In: Recent advances in engineering and computational sciences (RAECS)
15. Yu JY, Huang DF, Li Y, Zhang Y (2009) Implementation of MP3 player for music therapy on hypertension. In: 31st Annual international conference of the IEEE EMBS IEEE
16. Pingle YP, Anand B (2015) Music therapy and data mining using Indian Ragas as a supplementary medicine. In: 2nd International conference on computing for sustainable global development (INDIACom). IEEE, pp 347–350

17. Teng XF, Wong MYM, Zhang Y (2008) The long-term effect of music on hypertensive patients. In: Proceedings of the 5th international conference on information technology and application in biomedicine, in conjunction with The 2nd international symposium & Summer School on Biomedical and Health Engineering
18. Teng XF, Wong MYM, Zhang Y (2007) The effect of music on hypertensive patients. In: 29th Annual international conference of the IEEE engineering in Medicine and Biology Society Joint Research Center for Biomedical Engineering
19. Pillai A, Darshan D (2018) Evaluation of the effect of Indian classical music on the blood sugar level of type 2 diabetes mellitus patients. Music Med Interdisc J 10(4):175–179
20. Chavan D, Kumbhar MS, Rohit C (2016) The human stress recognition of brain, using music therapy. In: International conference on computation of power, energy, information and communication (ICCPEIC), pp 200–203
21. Nawsalkar R, Deshpande S, Butey P, Thakare V (2015) EEG based stress recognition system based on Indian classical music. In: International conference on advances in computer engineering and applications, pp 936–939
22. Kulkarni P, Rajesh J (2017) Impact of Raaga Malkauns on blood pressure measurement. In: International conference on computing, communication, control and automation (ICCUBEA), Pune, India. IEEE, pp 1–4
23. Leili T (2019) Multimodal machine learning for interactive mental health therapy. In: ICMI'19, Suzhou, China, pp 453–456
24. Chavan DR, Kumbhar MS, Rohit C (2016) The human stress recognition of brain, using music therapy. In: International conference on computation of power, energy information and communication (ICCPEIC). IEEE
25. Bardelar AA, Ajay G (2016) Study of Indian classical rags structure and its influence on human body for music therapy. In: 2nd International conference on applied and theoretical computing and communication technology (iCATccT 2016), Bangalore, India. IEEE, pp 119–123
26. Baharum A, Seong TW, Zain NHM, Yusop NMM, Omar M, Nordaliela Mohd R (2017) Releasing stress using music mood application: DeMuse. IEEE

Scottish Education: A Model for Strengthening Mental Health—A Case Study in Inclusive Scottish Primary Schools

Hanah Kim

Abstract Contemporary society represents a new wave of interests in Mental Health affected by the double-sided effect of development. A new wave of rapid technological development in human affairs has created greater opportunities for experiencing creativity in new forms that can then underpin enhanced states of human welfare. At the same time, these changes have become threats to human well-being, resulting from the dislocating impact of new ways of living. Despite the positive effects of improvement, negative consequences have arisen such as dehumanization due to the atomization of modern life, and the troubling spectacle of increased rates of Mental Health problems among the young (Mulraney et al. in Br J Psychiatry 217:364–369, 2019 [1]). Consequently, the role of schools for strengthening Mental Health has been increasingly emphasized. The concept of 'Mental Health' is especially meaningful to be investigated through Music Education since human beings have had an enduring relationship with sound and music. The principal aim of this paper is to investigate how schools can address issues of Mental Health and consider the potential of music-therapeutic interventions to be added in inclusive mainstream schools to expand the positive effects of music. First, in the conceptual part of the research, the benefits of music to Mental Health are discussed, particularly focusing on the concept of 'emotion'. Thereafter, the empirical part of this research is illustrated, which was conducted in Scottish primary schools to investigate how Mental Health is being approached. The analysis of data shows that Scottish schools are providing good models for addressing issues of Mental Health for a wide range of pupils from diverse backgrounds. This paper concludes that Music Education should continue to harbor rich possibilities for promoting Mental Health as an essential attribute of the flourishing and well-lived life.

This paper is an extracted and revised version of parts of the author's Ph.D. thesis: 'Creativity and Wellbeing in Music Education—philosophy, policy and practice in the context of contemporary Scottish Primary Education'. In this paper, the author focused on the concept of 'Mental Health', concerning the notion of 'Well-being'.

H. Kim (✉)
University of Glasgow, Glasgow G12 8QQ, UK
e-mail: therapist327@naver.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_28

Keywords Mental health · Emotion · Scottish education · Music education · Music-therapeutic interventions · Inclusive mainstream schools

1 Introduction

Contemporary society represents a new wave of interests in Mental Health issues. The term 'Mental Health' refers to a state of well-being in which individuals realize their own abilities to cope with the normal stresses of life and work productively to make contributions to their communities [2]. Ensuring the proper level of Mental Health is important to think, feel, and interact with each other to live and enjoy individual lives.

Due to the double-sided effect of development in the modern era, there have been increased occasions of the Mental Health crisis. Contemporary society has shown a new wave of development in human affairs. Resting upon rapid technological innovation, it has created greater opportunities for experiencing creativity in forms that can then underpin enhanced states of human welfare. At the same time, these changes have become threats to human well-being, resulting from the dislocating impact of new styles of living. Moreover, the eminent feature is that not only adults but also an increased number of young children and pupils are having problems with Mental Health [3]. As a result, 'Mental Health' has arisen as an important theme in the current era, chiefly as assets required for human beings: to live, to cope and to prosper.

Consequently, the role of schools for taking care of strengthening Mental Health for pupils not only in special schools but also in mainstream schools has been increasingly emphasized. Specifically, there have been increased interests in arts subject: as those are regarded as disciplines that especially have the potential to change mental states of pupils through expression and appreciation. The concept of 'Mental Health' is especially meaningful to be investigated through Music Education at present time, since human beings have had an enduring relationship with sound and music in almost all cultures on record.

In this paper, the theme of 'Mental Health' is examined including both conceptual and empirical research. First, the benefits of music to Mental Health is discussed, particularly focusing on the concept of 'emotion' and general and clinical influence of music. Thereafter, the empirical part of this research is introduced, which was conducted in Scottish primary schools to investigate how Mental Health is being approached and strengthened.

2 Conceptual Study: Benefits of Music to Mental Health

Benefits of music to Mental Health have been recognized since ancient Greece, particularly in relation to 'emotions', that shape the landscape of our mental lives. In

line with the family of viewpoints that admit and intuit a basic relationship between music, emotion and humanity, this tradition has always appreciated the potential of music to enhance health and well-being. Most eminently, it has been argued that music influences people's minds and the emotional life of listeners and music-makers. Furthermore, emotions are not just the fuel that powers the psychological mechanism of a reasoning creature but also parts of our cognitive and evaluative reasoning-selves too.

Since the period in which Monism appeared, in Wolff's *Logic* of 1728 [4], the dualism of Descartes' *Discours de la méthode* (1637) [5] has been challenged, with also David Hume (1748) emphasizing the 'embodiedness' of the human persons in which emotions are rooted and thereby strengthening the fundamentally Enlightenment belief that the human body and mind constitute one unitary entity [6]. Within this integrated understanding of their nature, human beings create and react to artwork through their body and mind [7], experiencing extended and extra-musical effects that are constantly interlayered with their health and well-being [8, 9]. The arguments were based on the presupposition that music is created by the expression of emotion and reveals emotion in different degrees and forms; and music affects emotion in terms of arousal and stimulation [10, 11].

Reflecting upon the effect of emotion in the life of human beings, the influence of music can be approached in two senses. In the first sense, the nature of emotional effects can be discussed in terms of health and well-being, which has seen a big increase in interest in today's society [12]—becoming indeed one of the major themes in contemporary education. Within this particular context, emphasis is currently placed on educational practice and a consistent perception of the curriculum upheld now in many parts of the world [13] that extends to topics such as emotional and Mental Health, and emotional resilience. In this currently powerful perspective, music is believed to be a source of emotional affect with the capacity to enlarge awareness and to move people who are involved with it [14]. As emotions are forms of evaluative judgment that attach themselves to certain things and certain persons outside the individual's own control, their cultivation is seemingly ratified across a spectrum of biological, cognitive, moral and social functions [15].

At the same time, the term emotion is also understood in other senses that are more directly therapeutic in character, signaling that music has an emotional value in dealing with, for example, the presence of damage or the problem of something that is going deeply wrong in the life of the individual and/or the community. In this perspective, music, it is frequently claimed, can help heal, and console, and make better—or make people more aware, or induce mindfulness and thoughtfulness, letting them experience the music they hear therapeutically. There is thus a strong emphasis today on the therapeutic, healing power of music in schools, where children often come from troubled backgrounds or carry particular challenges related to their emotional needs.

Music therapy is a professional practice in which sounds and music are used to address, support, and encourage the physical, emotional, cognitive, and social needs of individuals in more scientific and clinical ways [12, 16]. As Bruscia (1989) explains, music therapy is 'a systematic process of intervention wherein the therapist

helps the client to achieve health, using musical experiences and the relationships that develop through them as dynamic forces of change' [17, p. 47]. Discriminated from the generalized use of music for health and well-being discussed above, music therapy is a clinical and evidence-based use of music interventions to accomplish individualized goals within a therapeutic relationship by a credentialed professional [18]. A variety of clinical techniques and musical styles is used, depending on client needs [19].

The field of music therapy was established with the belief in the innate musicality of human beings and with the focus on the universality of music across cultures [20, 21]—on the assumption that music can positively influence health and well-being. Furthermore, a majority of music therapists insist that 'the ability to respond to music is not impaired by illness, injury, or disability' or 'limited by a condition or difficulty' [22]. Although the use of music for effects of this kind has been evident since the Greeks, the scientific application of music for such ends has become more pronounced since the Renaissance period, in which advances in medical science often combined key musical interventions on the basis of mathematical affinities and conceptions of universal properties such as 'harmony'.

Music therapists use the potential of musical affect and musical components to provide a means of relating within a therapeutic or rehabilitatory relationship. Their clients work with a range of accessible instruments and their voices to create a musical language which reflects their specific physical and emotional conditions, which then, it is suggested, further enables participants to meet their innate musical child—aiming 'to build connections with their inner selves and with others around them' [23] for the purposes of improving health and well-being. The benefits of music therapy then further extend to other disciplines, developing or improving physical, cognitive, social skills, and eventually building emotional resilience in clients, [24] which all influence states of 'Mental Health'.

3 Empirical Research: Mental Health in Educational Practice—Research Objectives, Methodologies, and Strategies

3.1 Research Objectives

The principal aim of this empirical part of the research was to investigate how schools can address issues of Mental Health and consider the potential of music-therapeutic interventions to be added in inclusive mainstream schools to expand the effects of music on Mental Health. Particularly, this research focused on Scottish Music Education. Scotland's *Curriculum for Excellence* (CfE) was designed in large measure to address the same growing emphasis on well-being. First, CfE makes considerable claims for well-being, emphasizing child-centeredness and autonomy—proposing

a process-based curriculum supporting an experience of well-being through interdisciplinary working. Secondly, CfE consists of the three main strands of literacy, numeracy, health, and well-being and aims for an inclusive vision of education that places the commitment to well-being at the center of its priorities. Furthermore, Scottish Education appears to be achieving these with a curriculum that is experimental, child-centered, and teacher-centered, and that allows for learners and teachers a great deal of autonomy and creativity. Above all, because it has a high-profile commitment to 'Health and Well-being', Music Education, seen to be in alliance with Mental Health, can be nourished in a number of different ways. Scotland has openly acknowledged the challenges of inequality, social exclusion, disadvantage, poverty, and attainment, and Music Education has responded by restating confidently its potential to address and handle these questions for all children.

Reflecting these unique features, it was expected that approaching Scottish primary schools and contacting experts would bring meaningful insights for the direction of education to address issues of enhancing Mental Health in more meaningful ways.

3.2 Research Methodologies and Strategies

The research herein consists of two types of interventions. The first part of the research was allocated to investigating and critically assessing the overall environment and atmosphere of the schools, and activities implemented in schools with the intention to enhance Mental Health. For this research, among a number of Social Scientific Research Methodologies, 'Ethnography' was chosen, with an intention to be fully embedded in the fields and to try proactively to enhance the health and well-being of children.

Ethnography is a broad term which describes the essentially dispositional features of naturalistic and interpretative methods [25]. It is a way of studying a group in a naturalistic context in fields often subject to long-term engagement with the research hosts [26, p. 120]. It aims at describing what happens in a particular work locale or social institution when it is in operation, particularly focusing on behaviors of the participants [27]. The author of this paper started working for embedded, 'ethnographic', and practitioner-enquiry investigations.

The author of this paper stayed in schools and participated in extra-curricular events to conduct participatory observations. Specifically, the aims for the observation were to consider the following questions, and the findings of each observed session were recorded in the Observation Proforma as the means to support research goals as represented in Table 1.

Thereafter, academic conversations with experts were conducted in order to examine deeper views of the philosophy and approaches in schools and to search for the directions that Music Education ought to follow in contemporary society to keep contributing on strengthening Mental Health. The themes included were as in Table 2.

Table 1 Observation proforma

Question themes
1. What was the learning intention of the session/program/event?
2. What kind of topics and contents did the instructor deal with in order to accomplish the learning intention?
3. How was this session/program/event organized? i.e., the order of the teaching and activities/program/event
4. What strategies did the instructor use to support the pupils' learning and participation?
5. How did the pupils interact with the instructor?
6. How did the pupils interact with other children?
7. How well were the pupils able to participate in the lesson/event and accomplish the learning intention?

Table 2 Themes for the interviewees

Question themes
1. What participants perceived as the fundamental benefits of music and music education
2. Participants' views as specialists of activities that are being used or ought to be implemented for fostering emotional well-being and emotional resilience

4 Research Places and Participants

To strengthen the practicality and attain more realistic reflections, research places were chosen among the Catholic primary schools located in deprived and disadvantaged areas where 'Mental Health' issues are taking eminent parts in education. Furthermore, interviewees were selected from multiple backgrounds and ages in order to situate the theme 'Mental Health' and view the practice of Music Education in wide and multi-dimensional ways. The pseudonymized names of the schools and the interviewees were as in Tables 3 and 4.

5 Results: How Schools Are Striving to Improve Mental Health

The results of this research showed that the concept of Mental Health was taking an outstanding place in school philosophy, policy, and curriculum.

First of all, all schools and interviewees believed in the power and benefits of music in terms of enhancing Mental Health. The research participants explained how music could be the medium for the expressions of feelings, and how it could change the 'inner mind', all resulting in the further consolidation of emotional capacity, also strengthening motivation and confidence. Several experts emphasized equally the

Table 3 Pseudonymized names of the schools and main features

School	Features
St. Magin's Primary School	An inclusive state school which specializes in hearing-impaired children supported by specialist staff. Championing inclusive education, the majority of classes consist of both non-disabled and disabled children
St. Petra's Primary School	A state school renowned in the local authority for its fine provision of Music Education in several different forms—achieved through continuous collaboration with universities and other musical institutions
St. Shelley's Primary School	A state school proclaimed as 'The School of Music'. The school puts a lot of effort into promoting Music Education with abundant musical resources and a lot of time allocated for musical activities and events, with a specialized music instructor being in a charge of running music sessions

Table 4 Pseudonymized names of the interviewees

Name	Position
Miss Wilson	A deputy headteacher in St. Petra's Primary School
Mrs Hay	A deputy headteacher in St. Shelley's Primary School
Miss Kelly	A classroom teacher in St. Petra's Primary School
Miss Ryan	A classroom teacher in St. Shelley's Primary School
Mrs Din	A classroom teacher in St. Shelley's Primary school
Mr Thomson	A music instructor in St. Shelley's Primary School
Mrs Foster	A music and religious education teacher in St. Magin's Primary School and a part-time lecturer in a local University which provides teacher education
Ms Winters	A musician and a music educator who has a great deal of teaching experience at all levels
Dr Murray	A singer-songwriter with abundant teaching experience in all different sectors with specialist academic expertise in history and in music
Mrs Hota	A Japanese sociologist who is running many musical performances and educational programs in schools for children

experience of being away from daily routines while they are involved in different forms of musical activity. Some interviewees stated as in Table 5.

In line with these beliefs, all schools and interviewees use music to improve Mental Health. Many programs and activities for cultivating emotional resilience were implemented in the chosen schools. Music in these contexts was playing an important role in enhancing pupils' mental well-being, both in mainstream and extra-curricular domains. There were common and unique activities that are implemented among different schools and teachers as indicated in Table 6.

Table 5 Statements of benefits of music for 'enhancing Mental Health'

Name	Statements
Mrs Din	I think it gives them an outlet, emotionally as well. For letting them express themselves…
Dr Murray	It touches the emotions in a way that is deeper than the word. And, if you can 'share' that with people, and 'they' realize it, and they will have love of that…. It can be a very 'healing' thing
Miss Wilson	I think it gives children opportunities beyond their normal, everyday lives. For example, children in this school have performed… So, you are taking them out of their normal place, and went have seen… maybe something that may inspire them. I think it is very good that… the children's confidence
Mrs Hay	So… for me, the music is like Yoga. It's for my… wellbeing. I feel I am in a different place…when I am singing or… being part of it. It takes you somewhere, and it [makes] you [get] away from stresses for a day. I find it very energizing

Table 6 Activities for strengthening Mental Health

Name	Activities
Miss Wilson	Singing songs related to 'emotion', Assembly and Masses including music-based activities
Miss Kelly	Meditation
Miss Ryan	Circle time, writing, meditation
Mrs Din	Games, scenarios, role-play, drama-making
Mrs Hota	Taiko, origami club, anti-nuclear movement, kimono corner
Mrs Hay	Circle time strategies, second step programme used across the school and in conjunction with the community work, emotional scaling type strategies, themed weeks across the school year e.g. anti-bullying week, health week, a focus on developing core values in children, i.e. the school values
Mrs Foster	Emotions in music, watch film, assembly including music-based activities, nurturing school, pot (nurture room), seasons for growth, rainbow room, teach rooms
Miss Winters	Playing instruments, 'The Big Noise'
Mr Thompson	Writing songs, recording, anti-racism/anti-beggary projects

As such, music was unanimously regarded across the expert group as an important source for educating pupils in broadly liberal and humanist terms and enhancing their Mental Health.

6 Analysis-Eminent Features and Theme: How Schools Are Striving to Improve Mental Health

As seen in the above data, it was possible to draw some themes and eminent features as the outcome of the research in the process of observing and listening to the opinions about how schools are striving to improve pupils' Mental Health.

First, musical interventions and activities were made across diverse subjects in the form of blended learning, addressing the philosophy and methods of *Curriculum for Excellence* (CfE). Placing music in different prisms enabled an increased number of pupils to experience the power of music for 'all', despite musical knowledge. This phenomenon can be interpreted with the concept of Gardner's multiple intelligence, as connecting music with different subjects could stimulate each pupil's different strength, set and guide journeys to be positively affected by music.

Secondly, it could be observed that the chosen schools were providing more expertised Music Education to increase the benefits of music to Mental Health. In all three schools, a block of timetables and special events were allocated to invite musicians, B.Ed. students, secondary school students to provide special lessons for improving Mental Health. Furthermore, in one of the inclusive schools for the hearing-impaired pupils, signing teachers were employed to support those pupils to learn and experience as other abled pupils. In this way, the theme 'Mental Health' was being addressed not only for pupils but also for the experts as well as they could develop leadership and realize the value of 'noblesse oblige'.

Finally, all these types of activities and the environment were enhancing confidence and self-esteem for pupils by providing them with opportunities to be active learners and learn to cooperate with others. For example, in the inclusive school for hearing-impaired pupils, both disabled and abled pupils had a set of classes to learn sign language and use those when performing which was a valuable way to nurture them to understand others with disabilities and difficulties.

Considering the philosophy of Scottish education and the disadvantaged environments where a large proportion of pupils continue to live, Music Education was particularly—in the distinctive context of confessional Scottish sectors—assuming a significant role in coordination with Religious Education. This ranged from the Religious Education curriculum itself to musically-enriched Roman Catholic Masses and other 'recreational-spiritual' events also designed to focus on nurturing pupils' personalities in settings spanning small groups to the whole school. Furthermore, devising extra-curricular activities such as school choir, allocating special sessions to gather all pupils in the school and be involved in special music lessons and events to let pupils be the performers were affecting pupils' Mental Health in positive ways.

7 Conclusion

As reviewed, the objective of strengthening 'Mental Health' is being relatively well addressed in Scottish primary schools, thanks to the philosophy and guidelines set out in CfE and in the curriculum's convincing appreciation of the benefits of music. Scottish schools were providing good models for addressing issues of Mental Health for *all*—encompassing a wide range of generations and people from diverse backgrounds. It could be observed that Scottish education system and CfE currently possess the potential to align Mental Health enormously effectively—chiefly by nurturing pupils who can respond creatively to our fast-changing world while fortifying their humanity and resilience. Disseminating the positive effects of music across the disciplines and co-working with professional and neighborhood communities including outer-school institutions, agencies of private, third and social enterprise sectors that would expand pupils' thoughts, minds, horizons, and actual experiences of music should be continued.

A number of researchers have stated that music is a 'non-discriminatory' way of engaging people [28]. The present research echoes this disposition in showing that carefully crafted music sessions can afford children opportunities to strengthen their Mental Health through the cultivation of empathy, self-expression and imagination in natural and empowering ways. What music can do in many different cultures, in its myriad uses of sound, is representing, describing, or narrating experiences that we have come to associate with certain events or emotions. There is no doubt that music can embody certain extra-musical associations or messages from our daily human lives [29] as well as from their moments of exceptional joy and exceptional sorrow. Therefore, music is indeed a medium for addressing Mental Health, that is among the significant drivers in the lives of all human beings everywhere.

Addressing the benefits of Music in practices will contribute decisively to the formation of pupils nurtured as capable members of contemporary society through the 'Four Capacities' that are aimed to achieve in Scottish education—to be Successful learners; Confident individuals; Responsible citizens; and Effective contributors. In line with the aims of cultivating 'whole persons' in the central stages of compulsory education, Music Education should continue to harbor-rich possibilities for realizing the core values of progressive education today: promoting Mental Health as an essential attribute of the flourishing, examined and well-lived life.

References

1. Mulraney M, Hiscock H, Sciberras E, Coghill D, Sawyer M (2019) Mental health difficulties across childhood and mental health service use: findings from a longitudinal population-based study. Br J Psychiatry 217(1):364–369
2. World Health Organisation (2018) Mental health: strengthening our response. World Health Organisation Homepage. https://www.who.int/mediacentre/factsheets/fs220/en/. Last accessed 2020/8/6

3. Lawrence D, Dawson V, Houghton S, Goodsell B, Sawyer MG (2019) Impact of mental disorders on attendance at school. Aust J Educ 63(1):5–21
4. Wolff C (2003) Logic (1728); Reprinted in logic, or rational thoughts on the powers of the human understanding. Olms, Hildesheim
5. Descartes R (1637) Discours de la méthode (Discourse on method). English version accessed from www.literaturepage.com/read/descartes-discourse-on-method.html. Last accessed 2020/8/6
6. Hume D (2007) An enquiry concerning human understanding and other writings. Cambridge University Press, Cambridge
7. Reybrouck M (2017) Expression and interaction in real-time listening: a dynamical and experiential approach to musical meaning. In: ESCOM 2017 Book of abstracts; 25th Anniversary Edition of the European Society for the Cognitive Sciences of Music (ESCOM); Expressive interaction with music, 31 July–4 Aug 2017, Ghent, Belgium
8. Hallam S (2015) The power of music. iMerc, University of London
9. Odena O (2018) Musical creativity revisited—educational foundations, practices and research. Routledge, London
10. Cohrdes C, Wrzus C, Riediger M (2017) Listening to self-selected music as mood regulation strategy? Interrelating regulation motivation, mood change and physiological reactivity in various ages. In: Conference paper; Proceedings of the ninth triennial conference of the European Society for the Cognitive Sciences of Music, 17–22 Aug 2015, Manchester, UK
11. Swaminathan S, Schellenberg EG (2015) Current emotion research in music psychology. Emot Rev 7(2):189–197
12. Brattico E (2017) Aesthetic empowerment through music. In: ESCOM 2017 Book of abstracts; 25th Anniversary edition of the European Society for the Cognitive Sciences of Music (ESCOM); Expressive interaction with music, 31 July–4 Aug 2017, Ghent, Belgium (2017).
13. MacDonald R, Kreutz G, Mitchell L (2012) What is music, health, and wellbeing and why is it important. In: MacDonald R, Kreutz G, Mitchell L (ed) Music, health, and wellbeing. Oxford University Press, Oxford, pp 3–11
14. Solberg R, Dobben N (2017) Break, build and drop: embodied listening as expressive interaction with electronic dance music. In: ESCOM 2017 Book of abstracts; 25th Anniversary edition of the European Society for the Cognitive Sciences of Music (ESCOM); Expressive interaction with music, 31 July–4 Aug 2017, Ghent, Belgium
15. Nussbaum M (2001) Upheavals of thought. Cambridge University Press, Cambridge
16. American Music Therapy Association homepage. https://www.musictherapy.org. Last accessed 2020/8/6
17. Bruscia KE (1989) Defining music therapy. Barcelona, Phoenixville, PA
18. Bunt L, Stige B (2014) Music therapy: an art beyond words. Routledge, London and New York
19. Lamont A, Knox R, Chau T, Hamdani Y, Schwellnus H, Tam C, Johnson P (2002) Converting movements to music—new musical exploration opportunities for children in rehabilitation. In: Canadian Association for Music Therapy 29th annual conference, pp 26–30, Regina, Canada
20. Mithen S (2005) The singing Neanderthals—the origins of music, language, mind and body. Weidenfeld & Nicolson, London
21. Blacking J (2000) How musical is man? University of Washington Press, Seattle, Wash
22. Arroyo-Anlló EM, Díaz JP, Gil R (2013) Familiar music as an enhancer of self-consciousness in patients with Alzheimer's disease. BioMed Res Int 1–10
23. British Association for Music Therapy homepage. https://www.bamt.org. Last accessed 2020/8/6
24. Thompson WF (2015) Music, thought, and feeling: understanding the psychology of music, 2nd edn. Oxford University Press, Oxford
25. Dey C (2017) Ethnography, ethnomethodology and anthropology studies in accounting. In: Hoque Z, Parker LD, Covaleski MA, Haynes K (eds) The Routledge companion to qualitative accounting research methods. Taylor & Francis, Milton Park, pp 147–162
26. Delamont S, Atkinson P (2018) The ethics of ethnography. In: Iphofen R, Tolich M (eds) The SAGE handbook of qualitative research ethics. SAGE, Thousand Oaks, pp 119–132

27. Hammersley M (2006) Ethnography: problems and prospects. Ethnography Educ 1(1):3–14
28. Odena O, Summers M, Jaap A, Rodríguez A (2016) Exploring the potential of music education for integrating newly arrived children in Scottish schools. In: Spoken paper. 32nd World conference, International Society for Music Education, Glasgow, 24–29 July 2016; In Music for integration—research briefing—exploring the potential of music education for integrating newly arrived children in schools, Robert Owen Centre for Educational Change
29. Stock JPJ (1996) Concepts of world music and their integration within western secondary music education. In: Spruce G (ed) Teaching music, Routledge in association with the Open University

Spoken Language Processing

Language Identification, Spoken Machine Translation, Speaker Recognition, Speech Transcription

Language Model Fine-Tuning with Second-Order Optimizer

Kartik Vyas and Dinesh Naik

Abstract At present, the artificial intelligence domain is completely dependent on first-order optimizers, the popular ones of which are stochastic gradient descent (SGD) and Adam. And it is obvious also as both of these give good rate of convergence and are lightweight as these have decent memory and computation requirements. This paper proposes a second-order training approach to train deep neural networks. The experiments are conducted in different domains, namely image classification and natural language processing. A comparison of first order and second order is presented for readers. In this study, the second-order quasi-Newton approach based on approximate Hessian calculation is applied to the famous natural language processing model Google BERT, which improves the peak accuracy in fewer iterations than Adam. This paper shows results of multiple implementations of second-order method on NLP and image classification tasks, and both these domains have shown improvements.

Keywords NLP: Natural Language Processing · SGD: Stochastic Gradient Descent · Adam: Adaptive Moment Estimation

1 Introduction

At present, the artificial intelligence domain is completely dependent on first-order optimizers, the popular ones of which are SGD and Adam [1]. And it is obvious also as both of these give good rate of convergence and are lightweight as these have decent memory and computation requirements.

A few characteristics which make first-order popular are:

- First-order methods move towards convergence in a single direction. That is, once they have a pivot, then they do not change course in normal scenarios and this direction is the direction of the steepest gradient, quite similar to the greedy algorithm paradigm which gives good results in most cases.

K. Vyas · D. Naik (✉)
National Institute of Technology Karnataka, Surathkal, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_29

- First-order gradients have no concept of maxima or minima, so in the initial iterations, it is impossible for these algorithms to calculate their relative distance from the final convergence point. This, however, is possible using a higher-order optimizer where the gradient of the step gives a better estimate of the local minima; however, finding a global minimum is still a far fetched objective.
- First-order methods require only the current on the functional curve. These do not take into account the steepness of the curve, which can only be found by double or higher derivatives. These only require the current loss to calculate the direction for the next iteration, therefore limiting the memory requirement.

The above-mentioned loop holes are exploited by higher-order methods. More specifically the second-order methods, where we are basically trying to locally model parabola simultaneously in multiple directions. There are mainly three challenges:

- Inverting Hessian: The first step in the higher-order optimization is calculating the Hessian matrix. Against the popular implementation of first-order methods, performing the calculations and storing the results for a second-order optimizer is $O(n^2)$ while storing the Hessian matrix in memory. This is huge because the mathematically equivalent equation for any deep learning model spreads in millions of dimensions and so has millions of parameters. For example, the Google BERT has 110 million parameters. There is some study done previously on second-order optimizer like in [2–4]. In all these, a modified version of Newton's algorithm is used. The modification generally is either truncation or an approximation. Preferring higher-order algorithms is only suitable if:
- Way more complex: Deep learning networks are in itself very complex as the number of layers and their interconnections have increased manifold, so if we introduce any more complexity we should expect to see relevant gains in performance.
- Handling saddles: Handling saddles is probably the biggest head ache of convex optimization field. Saddle points are basically points or planes where one set of dimensions have values showing increasing trends, and the rest shows a decreasing trend. This is a reason of concern because these planes consists of values which lead to a similar error rate, so escaping the saddle valley is difficult as there is no direction available to proceed with. We can escape saddles by performing multiple line searches so that before we actually enter it, we are already aware of it.

Despite many second-order methods have been proposed to train neural networks [5], most of the results do not compare results with Adam and no study has been done on BERT so nothing concrete can be concluded from these. This research work, however, removes that incompleteness by performing experiments on multiple tasks, namely NLP and image classification. And we compared the output of Adam and LBFGS.

2 Experimental Setup

The source code used to bind the datasets and the model architectures are written in Python 3.x as it is robust and has established itself in wide domains. Furthermore, the gradient calculation in which TensorFlow 2.x is used is an open source and free library. Table 1 shows the numbers of learnable parameters and layers in each model. The followings are the details of datasets and model architectures.

2.1 Model Architecture

2.1.1 CustomModel

It is a neural net with three convolution layers and three dense layers. The convolution layers have shapes of (30, 30, 32), (13, 13, 64), (4, 4, 128). There are two intermediate dense matrices and one classification dense layer, and three max pooling layers along with a flatten layer. The intermediate dense matrices have the output shapes of 256 and 128.

And the classification dense in the case of MNIST has the shape of 10, and Cactus dataset has the shape of 2 owing to its a binary classification task.

2.1.2 ResNet50

ResNet is short for residual networks and is a classic neural network used as a backbone for many computer vision tasks. It has 48 convolution layers, 1 maxpooling layer, and 1 average pooling layer. The ResNet50 model is pre-trained on ImageNet dataset so that fine-tuning is possible as the dataset for Cactus is limited and pre-training did not give any decent accuracy.

The reason for ResNet's popularity is that all its predecessors faced a very usual problem associated with every other neural network called the vanishing gradients problem, which is problematic because once the gradients vanish there is no learning in the model anymore but ResNet avoids it. Other architecture details can be found in the official paper [6].

Table 1 Table of model characters tics

Model name	Number of parameters	Number of layers
CustomModel	270 K	10
ResNet50	23.5 M	50
MobileNetV2	3.4 M	21
BERT	110 M	12

2.1.3 MobileNetV2

It is a slimmer model of MobileNet. It was primarily developed by Google to incorporate in mobile devices. MobileNetV2 is used in many mobile devices as its low-cost and lightweight API makes it favorable for mobile devices.

It has a first convolution layer with 32 filters and 19 bottleneck layers. These bottleneck layers take inputs and outputs from intermediate layers and translate them from low level features to high level features. And the last layer is the classification layer. The MobileNetV2 model is pretrained on ImageNet dataset so that fine-tuning is possible as the dataset for Cactus is limited. Other architecture details can be found in the official paper [7].

2.1.4 Google BERT

It is a machine learning-based model for natural language processing which has provided state-of-the-art results for many NLP problems like next sentence prediction, question and answering problem (SQuAD), masked token prediction.

BERT is an acronym for bidirectional encoder representations from transformers [8]. BERT, unlike other natural language processing models which parse the sentence sequentially to generate combination among words in one direction that is the way it is parsed, is not able to capture the semantics of the sentence completely that is they only analyze the sentence in one direction and so is able to understand only a partial meaning of the sentence. Whereas BERT reads the complete sentence and then generates a mutual embedding of each word corresponding to all other words both in the left and in the right to it. This helps the model to better understand the context of the sentence in a better way.

The BERT uses a transformer architecture [9], which uses the attention mechanism which captures the syntax and semantics very well, further discussion of which is out of scope for this thesis work. BERT has 12 attention heads, which implement the transformer architecture. Each of these attention heads has a feed forward network which basically does the task of learning three embeddings which will further result in learning the final word embed dings:

- query: this captures the task requested.
- key: the probable solution area the query leads to.
- value: the actual solution embedding.

The query matrix acts as a pointer to the Key:Value pair of the solution. The key value acts as a dictionary to the solution. These three matrices are learned by performing the dense operation with the input embedding.

2.2 Dataset

2.2.1 MNIST

MNIST [10] is an acronym for the modified national institute of standards and technology. This database is a subset of the NIST, a collection of labeled images of handwritten numerical characters or digits. The database contains black-and-white images of size 28×28 and is divided into a training and testing set of 60,000 and 10,000.

2.2.2 Cactus Dataset

Cactus dataset [11] is created by a team of Mexican researchers led by López-Jiménez in 2019, whose task is to list all endangered species of plants in Tehuacan-Cuicatlan Valley, which is marked as a heritage site by UNESCO. This team collected a set of 10,000 cactus images for training and 4000 for validation. The dataset consists of binary labeled 32×32 images where the labels indicate whether the image is cactus or not.

2.2.3 ImageNet

ImageNet is a crowdsourced dataset of over 14 million images distributed in 20,000 categories, and each category has several hundred images [12]. The project was initiated by Stanford's professor Fei-Fei Li in 2006, and it finished in 2009.

2.2.4 SQuAD

SQuAD [13] is commonly known as Stanford question and answering dataset which is assembled by Pranav Rajpurkar, who is a professor at Stanford. SQuAD is a reading and comprehension-based dataset which consists of 100 K sentences which are queries at Wikipedia by the general public. The task of the model is to predict whether the given question has an answer in the corresponding segment of the text provided.

3 Methodology

The two main steps in working of a second-order optimizer are calculating the hessian and, subsequently, it is inverse and the way these methods improve the line search direction using the exact or approximate of the previously calculated Hessian.

Using curvature information enables such methods to make more progress per step than first-order methods [14] relying solely on the gradient, more specifically talking about the Adam, the way it updates the variables is by learning the velocity and moment attributes. The update rules for velocity and momentum are then given as:

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g_t; \text{mean} \tag{1}$$

$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * g_t^2; \text{variance} \tag{2}$$

β_1 and β_2 are close to 1. The Adam update rule is thus:

$$\theta_{t+1} = \theta_t - \eta * \frac{m_t/(1 - \beta_1^t)}{\surd(v_t/(1 - \beta_2^t)) + \in} \tag{3}$$

Care should be taken to properly address the hyper parameters. On top of that as deep learning models are non-convex functions, we may or may not be going in the direction of reducing loss.

Newton updates, on the other hand, are solved by performing a Hessian vector product with the conjugate gradient (CG) inexactly [15], in which the search direction of the direction of the steepest slope is computed by applying CG to the Newton method. In the following paragraph, the Newton approach is explained.

3.1 Quasi-Newton Template

Let $x^0 \in R^n$, $B^0 \succ 0$. For $k = 1, 2, 3 \ldots$ repeat:

- Solve $B^{k-1}\Delta x^{k-1} = -\nabla f(x^{k-1})$
- update $x^k = x^{k-1} + t_k \delta x^{k-1}$
- compute B^k from B^{k-1}.

The way the new B_k is used to find the next step in the line search algorithm is:

$$\nabla f(x^k) = \nabla f(x^{k-1}) + B^k(x^k - x^{k-1}) \tag{4}$$

4 Results and Analysis

The experiments conducted on the models mentioned above were carried out a number of times by varying the parameters like the number of epochs, train and test data size, batch size, and so on. The setup, which resulted in the output closest to our expectations, is presented here.

4.1 CustomModel on MNIST LBFGS Versus Adam

The following experiment was conducted on the MNIST dataset and the model CustomModel. The experiment was run for 19 epochs. Both the loss Fig. 1 and the accuracy Fig. 2 plots show that LBFGS is superior in both aspects. One important observation from this is that the LBFGS curve reaches its peak or almost peak values in fewer iterations. This is as expected because, unlike Adam, a single search direction correction approach is implemented, the LBFGS does multiple corrections in the very short attempts. Also, the reason the Adam curve is recorded till the 9th iteration is because that is where we see the highest accuracy, after this point, the model overfits and no more improvements in performance are observed. These graphs also show that the impact of multiple iteration has on the convergence. For example, higher the value of 'iter' also known as iteration that is the number of line search conducted by LBFGS for each example, it corrects it's course multiple time leading to exceptional convergence. Although it requires more memory and time, those we will see in the conclusion section.

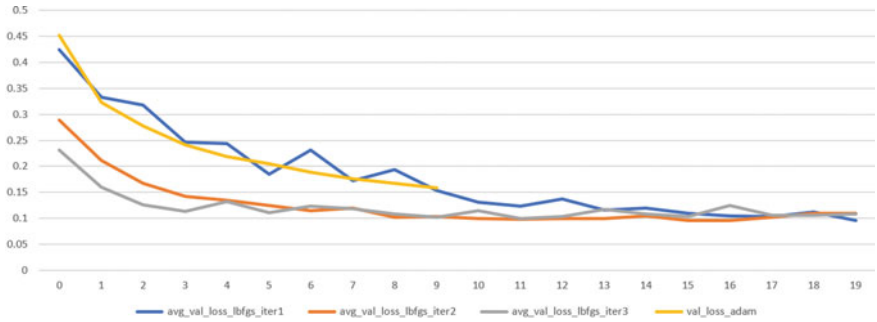


Fig. 1 Loss comparison between Adam and LBFGS

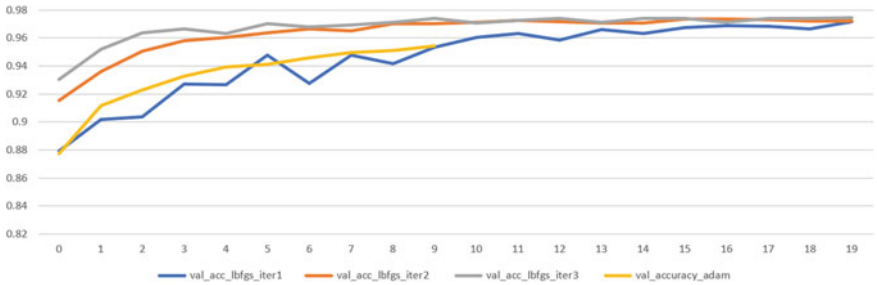


Fig. 2 Accuracy comparison between Adam and LBFGS

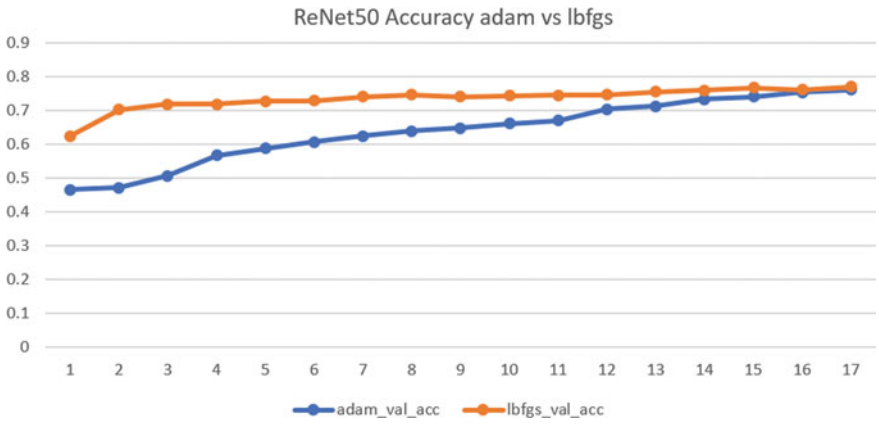


Fig. 3 Accuracy comparison between Adam and LBFGS

4.2 ResNeT50 on Cactus Dataset LBFGS Versus Adam

This is the accuracy graph Fig. 3 of Adam vs LBFGS on the Cactus dataset. The model is pretrained on ImageNet dataset, and fine-tuning is performed with Cactus dataset. The minor change is the architecture is the removal of the fully connected layers and replacing it with a separate classification layer. This is required to incorporate the difference in the number of classes in both the datasets. Here also we see that LBFGS reaches its peak accuracy value before Adam and that there is a large amount of improvement in the initial epochs and later the growth becomes linear with a small slope. This is where the Adam catches up with LBFGS. This experiment was conducted till 17 epochs as compared to the MobileNet version, where 10 epochs were required. This difference is due to the large number of trainable variables in ResNet50.

4.3 MobileNetV2 on CACTUS Dataset LBFGS Versus Adam

The accuracy comparison for optimizers LBFGS and Adam on MobileNetV2 model is presented in Fig. 4. The MobileNetV2 model was pretrained on ImageNet Dataset and fine-tuned on the Cactus dataset by removing fully connected layers from the original model and replacing it with a different classification layer, which has a separate fully connected layer to get new embeddings for Cactus dataset. The accuracy trend here also mimics the rest two experiments. The Adam tries to chase the peak accuracy already attained by LBFGS in much fewer steps. In comparison with ResNet50, we observe that MobileNetV2 requires fewer steps in the case of both Adam and LBFGS. This confirms the fact that steps required to train using any Adam or LBFGS is in some way related to the number of trainable variables in the

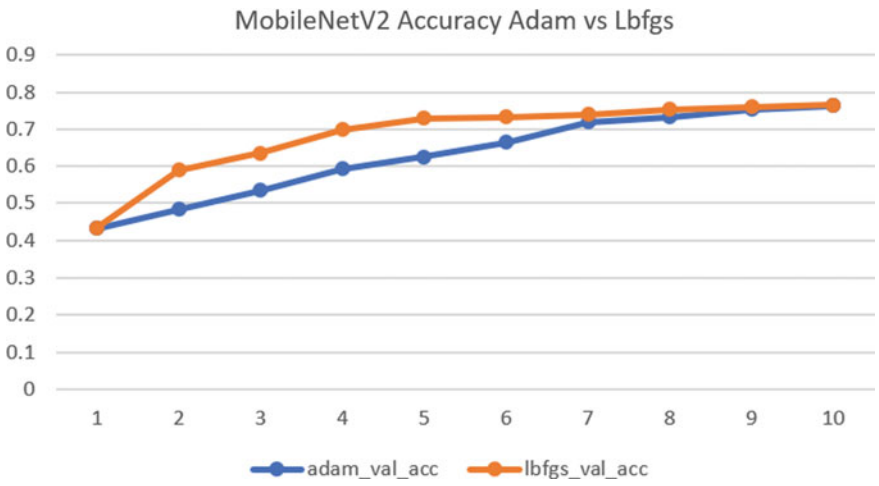


Fig. 4 Accuracy comparison between Adam and LBFGS

model. As there are less learn able features in MobileNetV2 as compared to ResNet50 it requires fewer steps to train, independent of the optimizer used. Again any further discussion on this domain will be done in the conclusion section.

4.4 BERT on SQuAD

The BERT model has produced metrics for the task of next sentence Prediction, where a pair of a sentence is input to the model, and the model predicts whether the sentences belong to the same context or not, thus making it a binary classification task. The official model available on GitHub was taken and fine-tuned for SQuAD. Google Research has provided a preprocessed dataset over SQuAD where from the context, two sentences are taken and labeled as next. Although a subset of that dataset was sufficient to train the model, we have already described the complete dataset in the above section. Figure 5 shows the accuracy comparison between the LBFGS and Adam, and Fig. 6 is the loss comparison. We can see that epoch six on the x-axis has shown better results for LBFGS.

5 Conclusion

We have performed different optimization tasks and on various models. Different tasks include both from domains like computer vision where images need to be classified in either 10 classes or binary classification or the task could be NLP where

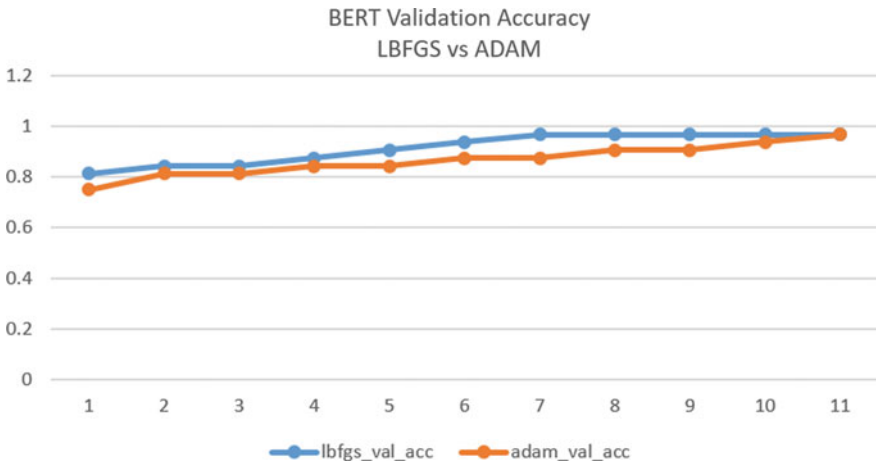


Fig. 5 Accuracy comparison between Adam and LBFGS

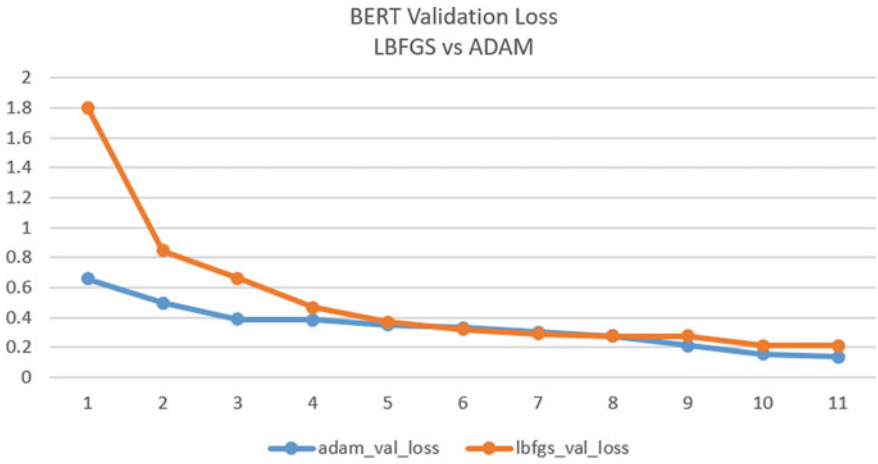


Fig. 6 Loss comparison between Adam and LBFGS

given a context, prediction needs to be made with respect the next sentence with similar context. All of which show that LBFGS is superior to Adam in terms of reaching the peak accuracy in fewer iterations. This is against the common belief where single-order optimizer is always suitable. Moreover, it can be seen that models with fewer parameters like the CustomModel, MobileNetV2 the LBFGS supersede the Adam by more margin, but in the case of ResNet50 and BERT, although LBFGS looks better, the difference is less. Further, for more iteration during line search leads to faster convergence though the memory requirement may also exceed. We need also remember that higher-order optimizer does not necessarily give better results in terms of the state-of-the-art models, but they help to reach those in fewer epochs.

6 Future Work

LBFGS has shown better results that Adam in terms of number of training steps required, still there is a trade-off with hardware requirement with time which needs to be studied further. Moreover, as there is very limited study in the domain of higher-order optimizer, without finding the limitations of higher-order optimizer the study is not complete. So understanding the strong ground and the limitations is still a distant study. We also need a comparative study of different higher-order optimizer that has been into practice. Most importantly, there has been a huge amount of study on distributed first-order optimizer like SGD but is it possible to distribute second-order methods in the same way as first order also needs to be studied. These are only some of the areas of research. Convex optimization holds a lot of scope for study.

References

1. Kingma DP, Adam BJ (2014) A method for stochastic optimization. https://arxiv.org/abs/1412.6980
2. Duda J (2019) Why second order SGD convergence methods are unpopular for deep learning? https://stats.stackexchange.com/questions/394083/why-second-order-sgd-convergence-methods-are-unpopular-for-deep-learning?newreg=f09d4fe13a044cabb2e08f361484ed7a
3. Taipale K Mathematical preparation for finance. https://www.softcover.io/read/bf34ea25/math_for_finance/multivariable_methods
4. Dobra A The newton Raphson algorithm for function optimization. https://www.stat.washington.edu/adobra/classes/536/Files/week1/newtonfull.pdf
5. Chen PH, Hsieh C (2018) A comparison of second-order methods for deep convolutional neural networks. https://openreview.net/forum?id=HJYoqzbC-
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
7. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) MobileNets efficient convolutional neural networks for mobile vision applications. https://arxiv.org/abs/1704.04861
8. Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. https://arxiv.org/abs/1706.03762
10. Cun L (210) Yann and cortes and corinna. MNIST: handwritten digit database. http://yann.lecun.com/exdb/mnist/
11. López-Jiménez E, Vasquez-Gomeza JI, Sanchez-Acevedo MA, Herrera-Lozada JC, Uriarte-Arcia AV (2019) Columnar cactus recognition in aerial images using a deep learning approach. Ecolog Informat 52:131–138. https://doi.org/10.1016/j.ecoinf.2019.05.005
12. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Miami, FL, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848

13. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ Questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing, pp 2383–2392. https://www.aclweb.org/anthology/D16-1264
14. Duchi J, Singer Y (2011) Adaptive sub gradient methods for online learning and stochastic optimization. http://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf
15. Bollapragada R, Byrd R, Nocedal J (2016) Exact and inexact subsampled newton methods for optimization. ArXiv preprint arXiv:1609.08502

Human-to-Robotic Voice Conversion by Employing Wave Patterns Based on Gender and Emotions

Dipankar Das and Rubel Das

Abstract Mutual conversion from one's voice into another is one of the promising research fields in speech processing communities while working with human voice. Influenced by the rapid interest in this area in recent trends, we have developed a system to identify gender and emotions, e.g., *sad*, *happy*, *angry*, *neutral*, and *fear* from human voice. A separate repository containing wave patterns of all possible phonemes of *male* and *female* voice in different emotions has been developed by employing this system. Similarly, we have prepared another system to convert human voice into equivalent phonemes. On the other hand, like human voice wave repository, an additional repository containing wave patterns of robotic voice has also been developed. Finally, the phonemes of human voice are converted into their corresponding phonemes of equivalent robotic voice based on the matched wave patterns available in this mapped repository. The results have shown satisfactory performance in voice conversion.

Keywords Human voice · Robotic voice · Speech · Emotion

1 Introduction

Speech is an information-rich signal that contains paralinguistic as well as linguistic information. As a result of this, speech conveys more emotional information than text. This reality motivates many researchers to consider speech signal as a quick, effective, and natural medium to identify interaction mysteries between computer and human. Although there is a significant improvement in speech recognition systems till date, still researchers are away from natural interplay between computer and human, since computer is not capable of understanding human emotional state. The recognition of emotional speech aims to recognize the emotional condition of individual utterances by applying his/her voice automatically. Recognizing of emotional conditions in speech signals is so challenging area for several reasons.

D. Das (✉) · R. Das
Computer Science and Engineering Department, Jadavpur University, Kolkata, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_30

Majority of the speech emotional methods used to select the best features that are powerful enough to distinguish between different emotions. The presence of various languages, accents, sentences, speaking styles and even speakers also adds another difficulty because these characteristics directly change most of the extracted features including pitch and energy [1]. Furthermore, it is possible to have a more than one specific emotion at a time in the same speech signal and each emotion may correlate with a different part of speech signals. Therefore, defining the boundaries between parts of emotion is very challenging task.

Automatic speech recognition and voice conversion became more useful in research to improve the human–machine interaction. One of the important concerns is to figure out how to incorporate the flavor of natural feelings into a human–machine interaction. In this regard, voice conversion can be served as a useful tool to provide new insights related to individuality of speech-enabled systems. However, voice conversion can be defined in different ways. It is considered as a method to modify the characteristics of speech given by a source speaker, so that it sounds as if it was spoken by a different target speaker [2]. On the other hand, it is an area of speech processing that deals with the process of hiding a speaker's identity [3]. Researchers also said that voice conversion is the adaptation of characteristics of a source speaker's voice to those of a target speaker [4].

More generally, the work in voice conversion can be extended to apply in many areas of speech processing like speech compression as well as in user-centric speech processing where only speaker is of main interest. It becomes essential to convert music of celebrities from one language to another language. Also, another important application of voice conversion is mimicking voice, preprocessing speech for recognition, restoring helium speech, etc. So far, a number of various approaches have been proposed for achieving voice conversion and most of them are based on non-parallel techniques [5]. However, collecting more corpora containing voices from both the source and target speakers and implementing parallel technique are considered as more efficient approaches. Based on the research results on speech individuality, it is generally accepted that voice conversion can be sufficiently achieved by comparing audio wave in phoneme level and modifying those of the source speaker into their corresponding phonemes of the target speaker [6].

In order to start with, we followed the similar approach. Our present approach also follows parallel training and mapping of every phoneme of a source speaker to those of its target speaker based on wave patterns. As in the case of a wave pattern, the frame energy and intensity of a voice may differ for different emotion types even for a single person. Such features also differ from person to person and therefore, the major challenge lies in identifying phonemes from the visualized wave pattern of an audio file. Thus, in our present attempt, we have developed a convenient approach to identify gender and emotion from an audio voice and utilized the information for further processing at phoneme level. In addition, we discussed our approach on how to identify phonemes of every utterance from an audio wave and finally how to convert the voice of a source to a target speaker.

The rest of the paper is organized as follows. Section 2 describes the related work done in this area. On the other hand, Sect. 3 narrates the information on dataset

or corpora used in the present task. Section 4 discusses on the first system that is developed to identify emotions and gender whereas Sect. 5 illustrates the strategy of second system on how to convert human voice into robotic voice by phoneme level mapping. Finally, Sect. 6 describes the experiments and results and Sect. 7 concludes with future works to be carried out in this area.

2 Related Work

If we observe a comprehensive review of speech emotion recognition systems targeting pattern recognition researchers who do not necessarily have a deep background in speech analysis, we notice three main aspects of this research field: (1) important design criteria of emotional speech corpora, (2) impact of speech features on the classification performance of speech emotion recognition (SER) and (3) classification systems employed in SER.

Multilevel support vector machine (SVM) classifier and artificial neural network have been used in [7] to reduce dimensionality by employing several parameters (e.g., energy, pitch, spectrum cutoff frequency, correlation density (Cd), fractal dimension, etc.) and obtained 86.5%, 68.5%, and 50.2% recognition rates at different levels on Beihang University Database of Emotional Speech (BHUDES). Similarly, the authors in [8] used binary classifier with prosodic and contour features to obtain 75.8% rate of recognition on SEMAINE functional data. In recent trends, SVM with prosodic and spectral features has been used in [9] and obtained 44.4% recognition rate on Berlin, LDC, and FAU Aibo dataset.

In addition to the above mentioned works, in [10], novel modulation spectral features (MSFs) for the recognition of human emotions in speech are presented. An auditory-inspired ST representation is acquired by deploying an auditory filterbank as well as a modulation filter bank, to perform spectral decomposition in the conventional acoustic frequency domain and in the modulation frequency domain, respectively.

The authors in [11] focused on the data preprocessing techniques which aim to extract the most effective acoustic features to improve the performance of the emotion recognition. The presented algorithm integrates the advantages from a decision tree method and the random forest ensemble. Experiment results on a series of Chinese emotional speech datasets indicate that the presented algorithm can achieve improved results on emotional recognition and outperform the commonly used principle component analysis (PCA)/multidimensional scaling (MDS) methods and ISOMap dimensionality reduction method.

In [12], a fusion-based approach to emotion recognition of affective speech using multiple classifiers with acoustic–prosodic information (AP) and semantic labels (SLs) is presented. The acoustic–prosodic information was adopted for emotion recognition using multiple classifiers and selects an appropriate classifier to output the recognition confidence. Moreover, in recent trends, so many related attempts on speech emotion recognition are available in literature [8, 13–15].

It is observed that all the above-mentioned approaches are either tried to deal with signals, acoustic features or to use machine learning classifiers and feature reduction techniques to improve the performance of speech emotion recognition (SER). In contrast, our proposed framework deals with both the emotion and gender of a speech together to capture the insights under the deep learning umbrella.

3 Dataset

Two systems, namely System 1 and System 2, have been developed to accomplish our research goals. In order to develop the module on gender detection with respect to System 1, we used 205 *female* and 210 *male* audio files for training purpose and all the audio files are in *.wav* format. The same training dataset has been used for the task of emotion identification as well. Besides, we used another 503 audio files (mixture of *female* and *male* voice) to test the performance of gender detection module. Similarly, we have added 177 and 258 audio files for *male* and *female* voices, respectively, to test the performance of emotion identification module.

In case of System 2, two types of dataset were taken, one containing audio files for *male* and *female* voice (human voice dataset) and another containing audio file of robotic voice. It is observed that human voice dataset is comparatively larger than robotic voice dataset. Thereafter, the human voice dataset is divided into two equal halves uniformly such that both halves have same diversity. One half is used to extract wave pattern of all possible phonemes of *male* and *female* voice in different emotions, separately and another is used as our test dataset. However, the full robotic voice dataset is used to extract wave patterns of all the phonemes of robotic voice.

4 System 1: Emotion-Gender Classification

One of the prime motivations of our first system, i.e., System 1 is to extract relevant features of an audio file and used them to find out emotion(s) and gender and of that audio. We have developed a classifier to classify emotion and gender of a human voice based on its important features. The followings are the key components of System 1: *Emotion Identifier* (feature extractor, classifier, and test module) and *Gender Predictor* (MFCC extractor, classifier, and test module).

4.1 Module 1: Emotion Identifier

In order to develop the module on identifying as well as classifying emotions, the first task is to extract relevant features.

Feature Extractor: This module is used to extract features of each audio file and hence, initially, it takes all the *.wav* files as audio input. The wave module of Python library[1] which provides convenient interface for the *.wav* sound format is used to read data from input file. It returns a tuple of two items, the sampling rate (i.e., samples per second) and an array data read from that file (an array containing a series of amplitude and frequency values which represent the input voice wave). We extracted the audio features, e.g., vectors, amplitude, frequency of input wave, respectively, and computed the root mean square (RMS) value of amplitude and average frequency and also the intensity of input wave. Finally, it will provide a vector containing three elements for each audio file. Output vector will be like <a, f, i> where a stands for RMS amplitude, f stands for average frequency, and i stands for intensity of the wave. The final vector has been produced by appending emotion value (1: *happy*, 2: *sad*, 3: *angry*, 4: *neutral*, and 5: *fear*) to each vector by denoting the emotion to which input audio voice belongs. This emotion value is externally provided as input only when training dataset is used to develop classifier.

Classifier: The output of feature extractor module is fed to the classifier. Initially, some random vectors (15–20 vectors, like 3–4 vector for each emotion class) are selected by employing *K-means* clustering algorithm and placed it into three-dimensional space by taking the first three elements of each vector into consideration. Then, the remaining training vectors are fed one by one to compute distance between centroid of existing clusters (only those clusters which have emotion value same as training vector) and those vectors. Cluster has been formed in such a way that the fourth value (emotion value) of all the vectors belonging to one cluster is to be same. We select the centroid which is the nearest and also its distance is less than a threshold value. The new vector is added to that cluster as well as centroid is recomputed. If the distance is greater than our threshold value, the new vector is treated as a new cluster. In this way, all such training vectors are plotted into the space to make different clusters and then it can be used as a classifier (Fig. 1).

Test Module: After developing clusters, it is used to identify emotion value of the audio files present in our test set. Like training, it also takes the test input vector in <a, f, i> format where a, f, and i stand for RMS amplitude, average frequency, and intensity of audio files, respectively. When a new test vector is arrived, it finds the points that are within a threshold and plotted into the space in a circular distance assuming that test vector as center of that circle. Then, we compute the distances between all those points and itself. The point which is the closest is assumed to be the most similar to our test vector and the emotion value of that corresponding point is assigned as the emotion value of the test vector.

Emotion identification results are shown in Table 1, and it shows emotions of how many *male* and *female* audio dataset are correctly identified. In Table 1, among 41 *male* audio files of *happy* emotion, 36 audio files are correctly identified. Similarly, among 52 *female* audio files of *sad* type, 39 audio files are correctly identified and so on.

[1] https://github.com/tyiannak/pyAudioAnalysis/wiki/3.-Feature-Extraction.

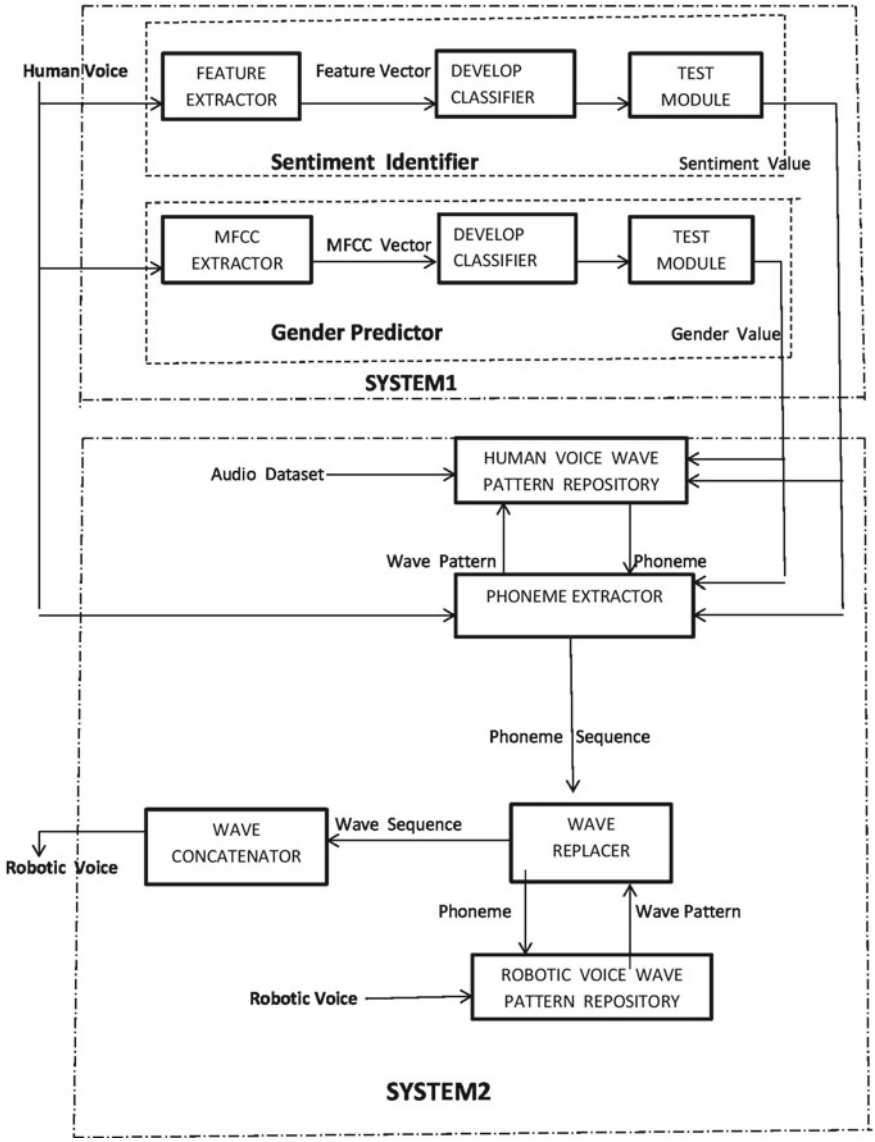


Fig. 1 Overall system architecture

4.2 Module 2: Gender Predictor

MFCC Extractor: This module takes audio files as input in *.wav* format and divides them into equal-sized frame. If the last frame is shorter, that will be padded with zero. Then, the following steps are followed to get MFCC vector.

Table 1 Emotion identification results of *male* and *female* audio dataset

Classes	Male			Female		
	Male	Female	Total[1]	Male	Female	Total[2]
Neutral	17	4	21	16	6	22
Happy	**36**	5	**41**	38	3	41
Sad	32	17	49	**39**	13	**52**
Angry	28	5	33	57	24	81
Fear	32	1	33	48	14	62
Total	**145**	**32**	**177**	**198**	**60**	**258**

1. Power spectrum value of each frame is computed and zero frame energy is replaced by epsilon value.
2. Calculate total energy, average frame energy, and filter-bank energies.
3. Then, zero filter-bank energies are replaced by epsilon value and output is a 2D array (*number of frames * number of filter*) where each cell contains energy value of each filter of each frame.
4. In order to normalize filter-bank energy value, we replace those values with its log value and compress filter-bank energies vector of each frame into shorter vector using discrete cosine transform (DCT) function.
5. Finally, the output is a 2D array with shape of (*number of frames * number of cepstral-coefficient*).
6. First, the coefficient of each vector is replaced by log value of frame energy.
7. Compute average cepstral-coefficient which is a vector for each audio input. As for example, if an audio input has k vector having m cepstral-coefficient each like $(x_{11}, x_{12}, \ldots, x_{1m}), \ldots, (x_{k1}, x_{k2}, \ldots, x_{km})$, the average coefficient vector will be

$$\left(\frac{\sum_{j=1}^{k} x_{j1}}{k}, \ldots, \frac{\sum_{j=1}^{k} x_{jm}}{k}\right) \tag{1}$$

Classifier: We build a classifier using MFCC vectors of training dataset. Similar to emotion classifier, this module also takes MFCC vector as input from the previous module. In order to make new vector of normalized value, we consider the first five coefficients for each of the vectors. Final vector is produced by appending one extra 4 value as 0 or 1 to it (0: *female*, 1: *male*). Initially, we take 10–15 random vectors and place it into five-dimensional space by taking first five elements of each vector into consideration. Then, remaining training vectors are come one by one and distance is computed between centroid of existing clusters (only those clusters which have same *gender* value) and those vectors. Cluster is to be formed such that the fourth value (i.e., *gender* value) of all vectors belonging to one cluster remains same. This gender value is provided as input at time of developing the classifier.

Table 2 Gender prediction results of *male* and *female* audio dataset

Classes	Male		
	Male	Female	Total
Male	***262***	55	317
Female	32	***154***	186
Total	294	209	503

Test Module: Once we develop the clusters, we focus to identify gender of our test audio files. It takes the test vector as input in <a_1, a_2, a_3, a_4, a_5> format which stands for normalized value of the first five MFCC coefficients of test audio file. When a new test vector comes, it finds some points plotted into space and the points are within a threshold of circular distance by assuming that test vector as center of that circle. Then, we compute distance between all those points and itself. The point which is the closest is assumed to be the most similar to test vector and gender value of that point is assumed to be the gender value of the test vector. Table 2 shows the gender of how many audio dataset is correctly identified. In Table 2, among 317 *male* audio files, 262 audio files are correctly identified and among 186 *female* audio files, 154 audio files are correctly identified.

5 System 2: Robotic Voice Conversion

As we discussed one of our main motivations regarding the development of System 2 is to extract wave patterns of all possible phonemes of *male* and *female* voices in different emotions and store them into a wave repository separately such that new wave pattern can be compared with those to find out which phoneme is represented by that new wave. In this section, we have discussed how to divide an audio wave into frames such that each frame represents one phoneme and also discussed how to compute similarity between two waves.

5.1 Human Voice Wave Repository

It acts as a repository to store wave patterns of every possible phoneme of *male* and *female* voices in different emotions, separately. Initially, the wave patterns of all unique single letter sound (like '*a*', '*z*', etc.) and then wave patterns of all unique sound generated by pair of letters (like '*ou*', '*ky*') are extracted from audio files of training dataset using Audacity tool.[2] Then, the extracted wave patterns are stored

[2]https://www.audacityteam.org/.

as map format[3] where phoneme symbol, gender value, emotion value together act as key and values are the corresponding wave pattern in *.wav* format. It means that it will store wave pattern separately for each phoneme, each emotion, and each gender value.

5.2 Robotic Voice Wave Repository

It acts as a repository to store wave pattern of every possible phonemes in robotic voice. Similar to human voice repository, here also, the wave patterns of all unique single letter sound and then wave patterns of all unique sound generated by pair of letters are extracted from robotic voice of the audio dataset using Audacity tool. Then, the extracted wave patterns are stored as map format where phoneme symbol acts as key and value will be corresponding wave in *.wav* format. It means that it will store wave pattern separately for each phoneme.

5.3 Phoneme Extractor

This module takes human voice audio file from test dataset as input. It also takes gender and emotion output values from *Emotion Identifier* and *Gender Predictor* modules of System 1, respectively. In our experiment, it was ideal to divide the entire wave patterns into 150–250 ms duration frames to get phoneme wise division (*a point between 150 and 250 ms, where the amplitude of wave is almost near to zero and is assumed to be end sound of one phoneme and starting sound of next phoneme*). Thus, initially, the wave patterns are divided into wavelets of 150–250 ms duration. After that, every wavelet is considered separately and compared with all human voice wave patterns stored in human voice repository to find out the most appropriate phoneme symbol for that wavelet (*compare with those wave patterns which have same emotion and gender value as input ones*).

It also uses cross-correlation technique to find out similarity between two waves. Similarly, the convolution technique can be used. More closer or similar wave pattern is taken as similar wave for that duration and corresponding phoneme symbol is considered for that duration. Finally, the output of this module is a sequence of phoneme symbols which represent that input audio. As for example, the wave patterns generated to pronounce word '*kite*' are considered as a sequence of phoneme symbol '*ky*' and '*it*' according to this system.

[3] https://github.com/vocobox/human-voice-dataset.

5.4 Wave Matching Implementation

Cross-correlation and convolution are closely related. To implement convolution with FFTs, the following steps have been followed:

1. Zero-pad the input signals (add zeros to the end so that at least half of the wave is '*blank*')
2. Take FET of both signals and multiply them element-wise.
3. Finally, we take the inverse of it. In Python, there is library function which is defined by,

$$\text{conv}(s_1, s_2) = \text{ifft}(\ \text{fft}(s_1_\text{and_zeros}) * \text{fft}(s_2_\text{and_zeros})) \tag{2}$$

In order to get cross-correlation, it needs to do time-reverse one of the signals before doing the FFT or take the complex conjugate of one of the signals after the FFT. In Python, there is library function which is defined by,

$$\text{corr}(s_1, s_2) = \text{ifft}(\text{fft}(s_1_\text{and_zeros}) * \text{fft}(s_2_\text{and_zeros[reversed]})) \tag{3}$$

$$\text{corr}(s_1, s_2) = \text{ifft}(\ \text{fft}(s_1_\text{and_zeros}) * \text{conj}(\text{fft}(s_2_\text{and_zeros}))) \tag{4}$$

Further, in order to get a similarity value between -1 and 1 (*a negative value indicates that if one of the signals decreases as the other increases*) it is needed to provide the amplitude corresponding to the length of the inputs, length of the FFT.

5.5 Wave Replacer

It takes a sequence of phoneme symbols as input from Phoneme Extraction module and passes all symbols one by one to robotic voice repository to get the wave patterns corresponding to those phoneme symbols. Finally, the output of this module is a sequence of wavelets of robotic voice corresponding to input sequence of phoneme symbols.

5.6 Wave Concatenation

It takes a sequence of wavelets from wave replacer module and concatenates sequentially one after another to generate final robotic voice which has same utterance as input audio file but voice is of robotic sound (Fig. 2).

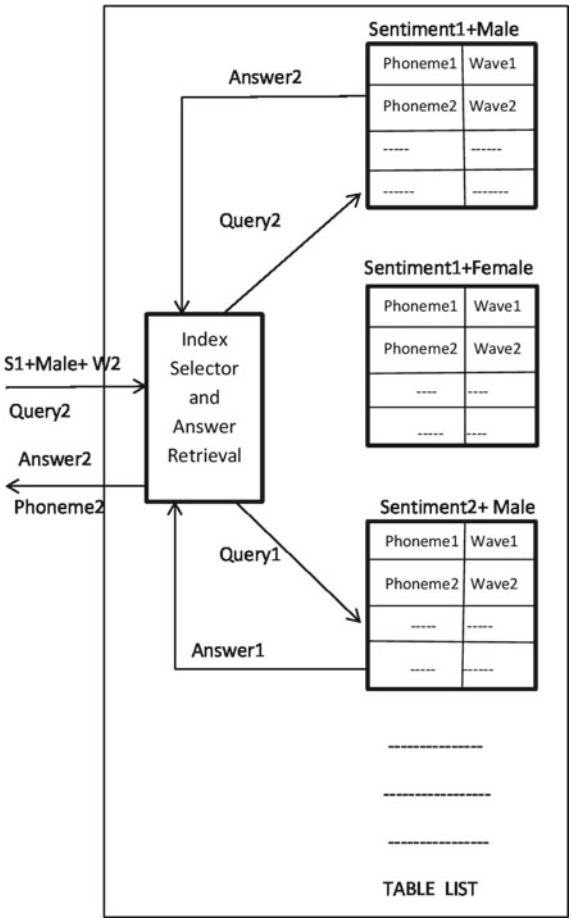


Fig. 2 Wave repository architecture

6 Results

It has to be mentioned that the evaluation of human to robotic voice requires an annotated dataset of emotion as well as robotic types from both the ends. To the best of our knowledge, such parallel voice dataset is not open source so as for the evaluation purpose. Therefore, we proposed new metrics for evaluating the performance of our systems.

For System 1, we used 177 *male* and 258 *female* audio files with different emotions to test the performance of *Emotion Identifier* module and also used 503 audio (*male* and *female* both) files to test the performance of *Gender Predictor* module. Among 177 *male* audio files with different emotions, 145 audio files are correctly identified (as shown in Table 1). Similarly, among 258 *female* audio files, emotions of 198 audio files are correctly identified (as shown in Table 1). So this module correctly

Table 3 Results on system 2

Rating	Percentage of file (%)
5: Very good	20
4: Good	33.33
3: Average	26.7
2: Below average	13.3
1: Non-acceptable	6.67

identified emotions of 343 out of 435 audio files, i.e., it achieved an overall accuracy of 78.85%.

On the other hand, among 503 audio files, gender of 416 audio files is correctly identified (as shown in Table 2), i.e., gender predictor module achieved 82.7% accuracy. For System 2, we proposed 1–5 rating scheme (5: *very good*, 4: *good*, 3: *average*, 2: *below average*, and 1: *non-acceptable*) to evaluate depending on the quality of generated robotic voice audio. Table 3 represents the results of System 2. If we look in Table 3, only 20% generated robotic voice audio are of *very good*, 33.33% are *good,* and so on. Overall, 80.03% of the generated robotic voice audio is satisfactory.

7 Conclusions

In this paper, we described how to initially extract emotions and gender information from audio file and also to develop a repository containing wave pattern for each phoneme in each emotion for both *male* and *female* voice as well as their corresponding robotic voice. Then, we explained our approach how to convert from human voice to robotic voice keeping same utterance. Non-parallel training of voice conversion can be another technique but parallel technique is better as it can do conversion of voice more accurately and efficiently. Here, the success of our system depends on resource and diversity of training dataset as if training dataset is more diverse and resourceful, it makes human and robotic voice repository (mentioned in System 2) more resourceful and helps to implement efficient and accurate mapping between the voices of source and target speakers.

Acknowledgements The work reported here is supported by the grant in aid from the IMPRINT-II sponsored research project, 'Sevak—An Intelligent Indian Language Chatbot', funded by SERB, Government of India.

References

1. Russell SJ, Norvig P, Canny JF, Malik JM, Edwards DD (2003) Artificial intelligence: a modern approach, vol 2. Prentice Hall, Upper Saddle River
2. Moulines E, Sagisaka Y (1995) Voice conversion: state of the art and perspectives. Speech Commun 16(2)
3. Stylianou Y, Capp O, Moulines E (1998) Continuous probabilistic transform for voice conversion. IEEE Trans Speech Audio Process 6(2)
4. Acero A (1998) Source-filter models for time-scale pitch-scale modification of speech. In: Proceedings of the ICASSP98, Seattle, USA
5. Myers CS, Rabiner LR (1981) A comparative study of several dynamic time-warping algorithms for connected word recognition. Bell Syst Tech J 60(7)
6. Sndermann D, Ney H, Hge H (2003) VTLN-based cross-language voice conversion. In: Proceedings of the ASRU03, Virgin Islands, USA
7. Chen L, Mao X, Xue Y, Cheng LL (2012) Speech emotion recognition: features and classification models. Digit Signal Process 22(6):1154–1160
8. Arias JP, Busso C, Yoma NB (2014) Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Comput Speech Lang 28(1):278–294
9. Cao H, Verma R, Nenkova A (2015) Speaker-sensitive emotion recognition via ranking: studies on acted and spontaneous speech. Comput Speech Lang 28(1):186–202
10. Wu S, Falk TH, Chan W-Y (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53(5):768–785
11. Rong J, Li G, Chen Y-PP (2009) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag 45(3):315–328
12. Wu C-H, Liang W-B (2011) Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans Affect Comput 2(1):10–21
13. Basharirad B, Moradhaseli M (2017) Speech emotion recognition methods: a literature review. In: Proceedings of AIP conference
14. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA (2018) Speech emotion recognition: methods and cases study. In: Proceedings of the 10th international conference on agents and artificial intelligence, pp. 175–182
15. Frant E, Ispas I, Dragonir V, Dascalu M, Zoltan E, Stoica LC (2017) Voice based emotion recognition with convolutional neural networks for companion robots. RJIST 20(3):222–240

Classifying Hate Speeches Shared in Twitter

Dipankar Das, Souvik Mondal, and Antariksha Ray

Abstract The growing use of online social networks and numerous interactions also lead to a marked growth in generating hateful contents across mass media. Thus, it has opened a venue to analyze hate speeches appearing online and develop systems to filter out such contents before recommending them to the viewers. In the present work, we have attempted to prepare a comparative study regarding the usages of different features like *n-gram*, *sentiment*, *sentence vectors*, and their combinations to detect and classify hate speeches by employing them in a neural network. Our approach not only extracts the above features from tweets but trains the neural network also based on these features. We identified the best model while using *4-gram* word features with *sentiment,* and the model achieves an average accuracy of 73.6%. We also performed some additional experiments and observed that when the training data is equally distributed, neutral and hate speeches are classified with very high accuracy. Similarly, we found that the classification accuracy is improved when the training data contains a combination of the best individually performing features.

Keywords Hate speech · Tweet · Classification · Neural network

1 Introduction

Hate speech is a type of speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. With massive explosion of online social interactions, hateful contents are also increasing significantly especially in twitter [1]. This motivates us to automate the detection and analysis of hate speeches from tweets (e.g., "*@AndreBerto word is you use roids, stupid hypocrite lying faggot*"). Moreover, such a system can be used to identify and track public sentiment as well as to segregate a group of users from another group or filter out such hateful content, as required. There are always confusions regarding the types or classes of hate speeches as the degree of their impacts is difficult to be judged.

D. Das (✉) · S. Mondal · A. Ray
Computer Science and Engineering Department, Jadavpur University, Kolkata, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_31

In our present work, we have classified hate tweets into three types—"*neutral*", "*offensive but not hate speech*", and "*hate speech*". An *offensive* tweet may not necessarily be *hateful* toward anything in particular but a *hateful* tweet is considered as *offensive* (e.g., "*Warning: penny boards will make you a faggot*"). On the other hand, a *neutral* tweet is neither *offensive* nor *hateful* ("*@NFLfantasy @AkbarGbaja the last thing he cares about is fantasy owners*").

Challenges are observed during classification as *n-gram* features are mainly responsible in misclassifying *offensive* as *hate speech*. Many of the tweets considered most *hateful* contain multiple racial and homophobic slurs, meaning that we are likely to misclassify *hate speech* if it does not contain any curse words or *offensive* terms. More accurate classification of such cases can be performed if we find the sources of our training data that are *hateful* without necessarily using particular keywords or *offensive* language. Tweets containing dark *humor* are often treated as *offensive* in nature even though they may not have been originally posted with such intent. Our experiments were performed on English tweets only, so a different language could pose new challenges. In our experiments, we found that the significance of features is in this order (from the least to the most): *sentiment* < *sentence2vector* < *4-gram* < *all unique words* < *4-gram with sentiment*. Finally, we observed that while taking the *4-gram* with *sentiment* featured model and equally distributed training data, the model outperforms the rest.

The rest of the paper is organized in the following manner. Section 2 briefly discusses the related work carried out in different languages like English, Indian, and Chinese available till date. Section 3 describes the overall system framework containing different modules based on divided datasets along with feature studies. Modulewise experiments and results are discussed in Sect. 4 along with various case studies. Section 5 presents the experiments with detailed error analysis of results. Finally, conclusions and future directions are presented in Sect. 6.

2 Related Work

The task of identifying *hate speech* and abusive language in online forums is being researched for more than 20 years [2]. In this survey, the authors pointed out what are the wings that have expanded in recent trends related to hate speeches so far. However, there are a few works that have been conducted in the case of Indian languages. In connection to social networks and the usage of code-mixed texts, researchers engaged themselves to show the potential of hate speeches in the context of Indian languages by exploring deep neural architectures [3, 4]. A different combination of CNN, LSTM, TF-IDF, GloVe embedding, and random embedding was used in [3] for a high F-score of 0.93. [5] used CNN for classifying twitter text using n-gram, and word2vec features with a 78% F-score. On the other hand, *offensive* language topic clusters were developed using logistic regression over a set of 860,071 tweets automatically annotated using a boot-strapping technique and supplemented with a dictionary of 339 *offensive* words [6]. In contrast to ours, in the work described in

[7], the authors used unigram, bigram, trigram features with weighted TF-IDF and logistic regression with a $F1$-score of 0.9. Bag-of-words approaches generally have high recall but also have high false positives since the presence of offensive words can lead to the misclassification of tweets as *hate speech* [8]; syntactic features are proved to be better to identify the targets and intensity of *hate speech* in [9].

3 System Design

In order to obtain automated classification of *hate* tweets, we use a convolutional neural network (CNN) with three layers of input, hidden layer with 31 units, and output layer. At first, character *4-gram* features are extracted from the raw data and such features are used to generate the training matrix as well as train and test of the neural network.

3.1 Feature Vector Preparation

In order to identify the polarity of the tweets to measure the potentiality of a tweet to be of type *hate*, the tweets were assigned with an additional sentiment score by calculating the cumulative sentiment scores of the words extracted from the SentiWordNet.[1] It was observed that the SentiWordNet is unable to handle multiple query words and thus, we preprocessed the tweets in the following way, (1) tokenization, (2) part-of-speech tagging, (3) stopwords removal, and (4) normalization using lemmatization by nltk toolkit.[2] The tweets were again fed into the SentiWordNet. Multiple sentiment scores were assigned by finding the synsets of the individual words and then taking average of the scores of each word to find the score of the entire tweet. The different variations of sentiment scores used are: (1) *positive* scores, (2) *negative* scores, (3) difference of the *positive* and *negative* scores, (4) difference of the maximum *positive* and *negative* scores, and (5) difference of the minimum *positive* and *negative* scores. Finally, a sentiment feature vector is generated in the following way: [original class tweet] and the corresponding feature vector is *<avg(pos-neg), min(pos-neg), max(pos-neg), avg(positive), avg(negative), max(positive), max(negative), min(positive), min(negative), pos_score(SentiWordNet), and neg_score(SentiWordNet)>*. For example, the sentence "*2 warning penny boards will make you a faggot*" will produce the following scores for a feature vector.

−0.01, −0.15, 0.25, 0.029, 0.039, 0.25, 0.2, 0.0, 0.0, 0.0, 0.0.

[1] https://sentiwordnet.isti.cnr.it.

[2] https://www.nltk.org/.

4 Experiments and Results

We obtained 16k+ annotated tweets available from the crowdflower[3] dataset. After preprocessing the dataset, we selected 12,088 tweets from it, and changed all digits present in the tweets to "1" so the total number of features has been reduced. Out of all the remaining tweets, 5918 tweets are labeled as "*neither offensive nor hate speech*" (i.e., Class 1 in our case), 3978 are labeled as "*offensive but not hate speech*" (Class 2) and 2192 tweets are labeled as "*hate speech*" (Class 3). However, the distribution of different classes in the full dataset is not equal as shown in Fig. 2. Thus, we divided the total number of tweets into five separate datasets (D1 to D5) for conducting various experiments to identify the important features responsible to detect hate speeches and other classes too. Moreover, we have developed five different modules for case studies and have shown performances on all these five variants. The *Dataset6* was prepared with equal distribution of tweets from the three classes (300 tweets of each class, 900 total tweets in the training dataset) (Fig. 1).

4.1 Module 1: Ideal Case with All Unique Words

In order to build this module, all tweets were converted into lowercase and unrecognizable characters were removed. All the unique words were taken as the feature set. The overall accuracy of D2 is 70.89%, D3 is 72.42%, D4 is 71.43%, and D5 is 71.85%. It forms a baseline for classification.

4.2 Module 2: 4-Gram

All tweets were converted to lowercase, stemmed[4] and unrecognizable characters were removed. We extracted relevant words from all documents using TF-IDF. Some words with low scores that appeared across most of the documents were also selected if they contributed toward *hate* speech detection (e.g., "*bitch*", "*faggot*", etc.). A list of *offensive* words was scanned for this verification. For reduction of the character *4-gram* features, we eliminated all features that obtain less than a benchmark score (obtained by TF-IDF weighting of the features from the datasets) but not arising from any *offensive* word. Ratio for training and test dataset was 20:80, i.e., we trained the neural network with *Dataset6* over 100 epochs to provide equally distributed information for all three classes. The overall accuracy of D2 is 71.68%, D3 is 69.72%, D4 is 68.94%, and D5 is 70.41%. It appears that the performance has dropped very slightly as compared to the previous module (as shown in Fig. 3).

[3] https://data.world/crowdflower/hate-speech-identification.

[4] https://www.nltk.org/modules/nltk/stem/porter.html.

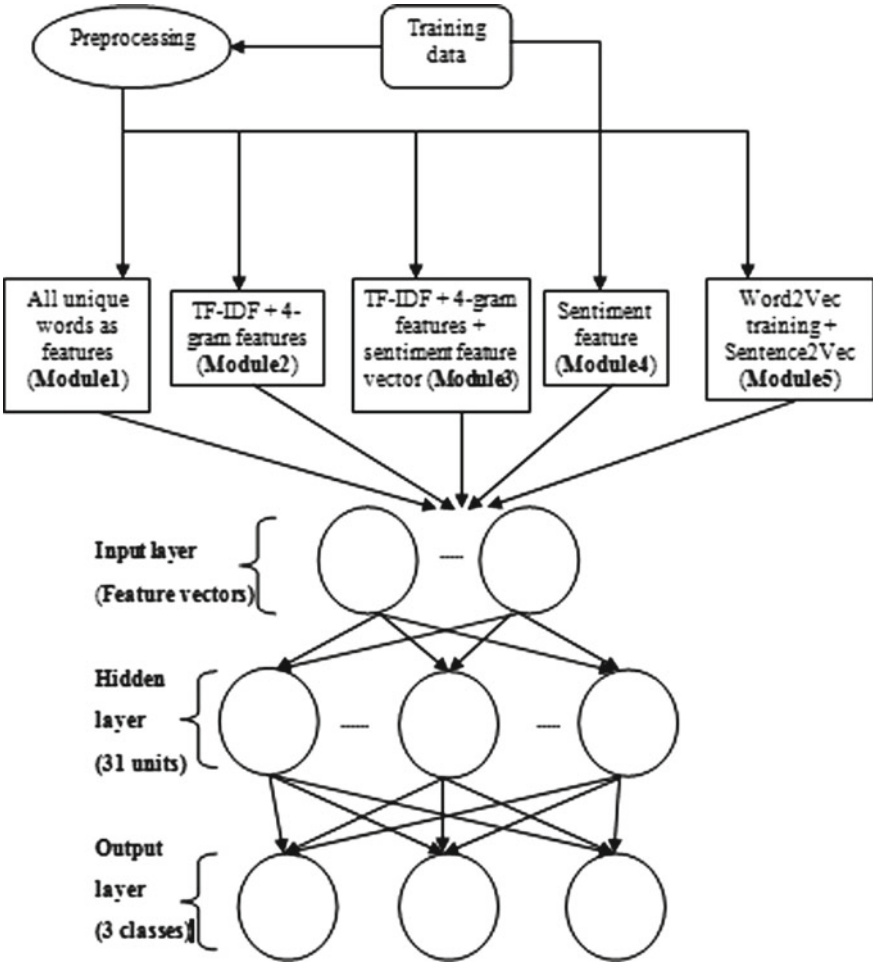


Fig. 1 Block diagram of CNN with features

4.3 Module 3: 4-Gram with Sentiment

In addition to the feature set from Module 2, the sentiment vector as a feature was added to train the network. We have shown the fully expanded results for our best case. Along the row, we have Class *i* and along the columns, we have the number of tweets that have been predicted as belonging to Class *j*, for datasets (D2 to D5). A column graph for each module shows the results graphically, e.g., Dataset2 (D2) of Class1: *Neutral* in the graph means *Neutral* class of Dataset2 has around 89% tweets predicted correctly as *Neutral*, which can be checked against Table 1 (as 1045/(1045 + 88 + 40)). The overall accuracies for D2 to D5 are 76.58%, 72.71%, 72.43%, and

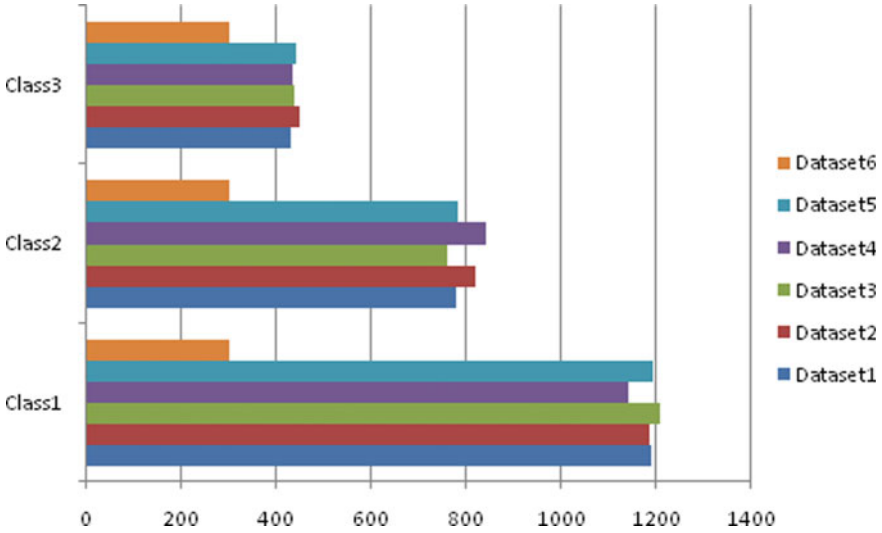


Fig. 2 Classwise distributions of the full data

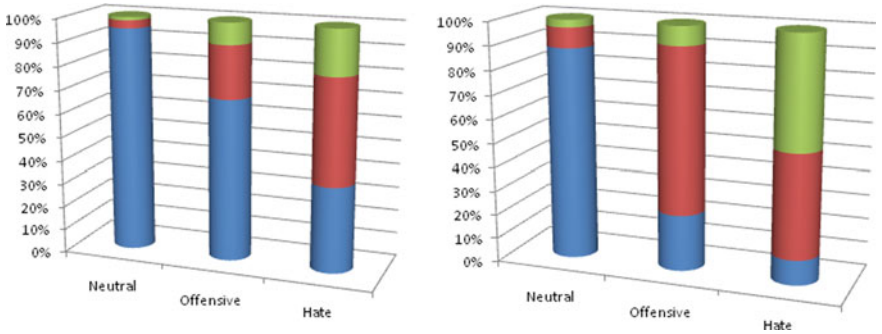


Fig. 3 Classification accuracies (in %) of module 1 and module 2

Table 1 Overall confusion matrix for module 3 w.r.t classes versus datasets

	Neutral (predicted)				Offensive (predicted)				Hate (predicted)				Total (predicted)			Total
Dataset (Accuracy)	D2	D3	D4	D5	D2	D3	D4	D5	D2	D3	D4	D5	Neutral	Offensive	Hate	
Neutral	1045	1073	1000	1054	88	105	88	93	40	35	42	36	4172	374	153	**4699**
Offensive	90	97	107	89	447	411	436	389	284	269	274	276	383	1683	1103	**3169**
Hate	18	48	38	36	51	107	109	119	376	278	293	278	140	386	1225	**1751**
Total	**1153**	**1218**	**1145**	**1179**	**586**	**623**	**633**	**601**	**700**	**582**	**609**	**590**	**4695**	**2443**	**2481**	**9619**

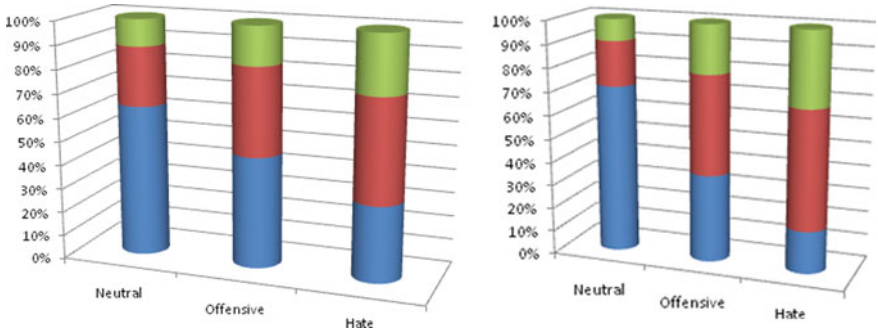


Fig. 4 Overall accuracies (in %) for module 3 and module 4

72.61%, respectively. This has provided an improved performance as compared to the previous modules.

4.4 Module 4: Only Sentiment

The tweets were converted to lowercase. The sentiment vectors for the tweets were used as feature vectors to train and test the neural network. Overall the accuracy of D2 is 46.98%, D3 is 45.43%, D4 is 45.10%, and D5 is 45.45%. It was observed that the performance strongly indicates that sentiment scores by themselves alone are not useful at all for this classification task (as shown in Fig. 4).

4.5 Module 5: Sentence2Vector

Similar to other modules, the tweets were converted to lowercase. Word2Vectormodel was trained on the dataset and then the vector of each tweet was calculated as the average of word vectors. This sentence vector was used as the feature to train and test the neural network. Overall accuracies on D2 to D5 are 53.82%, 55.46%, 52.18%, and 54.89%, respectively. The performance is slightly higher than random classification, which is not desirable and so this approach is not useful just by itself (as shown in Fig. 5).

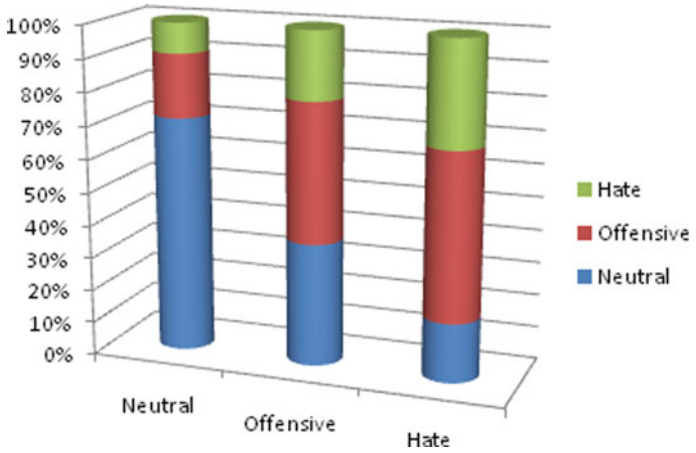


Fig. 5 Accuracy (in %) for module 5

4.6 Case Studies on Module 3

It was observed that Module 3 outperforms all the other modules in terms of overall as well as individual class level accuracies. Therefore, we planned to conduct detail experiments based on different cases as follows.

Case 1: 1400 tweets from each of the classes were combined into a single dataset and this dataset has been used to train our neural network and tested on the remaining tweets.

Case 2: Here, the dataset is divided into two parts, training (1428 * 3) tweets and test (700 * 3) tweets (3: *for all three classes*), such that the two sets have an equal number of tweets from each class. The neural network is trained and tested on these sets.

Case 3: The dataset was trained unequally using the set of all tweets except the test tweets from *Case 2*. The final classification is carried out by giving preference to *Case2* in case of *hateful* tweets. Otherwise, the preference is given to *Case 3*.

Case 4: Final classification is done based on the maximum class probability among all the six probabilities assigned by the network when trained unequally and equally as in the above two cases.

Case 5: The datasets D2 and D5 are combined and used to train the neural network and then the rest of the datasets were used for testing. This is done since D2 and D5 were generally exhibiting the highest accuracies in testing in the previous modules.

Case 6: The dataset is trained as in *Case2* but tweets having tokens from exception list = ["*fagg*", "*faggot*", "*nigger*", "*nigga*", "*fag*", "*faggots*", "*niggers*"] were not considered since these features lead to most misclassification between the classes,

offensive but not hate speech and *hate speech*. The confusion matrices (as shown in Table 3) and the results with respect to all the case studies are provided in Table 2.

5 Error Analysis and Observations

The tables below show the classification accuracy scores obtained on the four datasets (D2 to D5) after training the neural network. Along the row we have $Class_i$, and along the columns, we have the number of tweets that have been predicted as belonging to $Class_j$, e.g., 4231 *Neutral* tweets have been correctly classified in Module 1.

The performance drop in Module 2 as compared to Module 1 is due to the fact that although hateful tweets are much more correctly classified in Module 2, it is outweighed by the effect of increased misclassification of *neutral* and *offensive* tweets. However, it is observed that in Module 3, the classifier performs better overall by a margin of 3–4% when trained with Dataset6 and sentiment score. It is able to increase the correctness in the classification of Class 1 (*neither offensive nor hate speech*) and Class 2 (*offensive but not hate speech*) but drops very slightly in accuracy for Class 3 (*hate speech*). However, the overall classification accuracy has increased.

Limitations in Module 4 come from the way SentiWordNet provides multiple scores for a word. Based on different contexts, a word can have a *positive* or *negative* polarity. Capturing the context of the word would enable better sentiment scoring, e.g., "*You are extremely dirty. I am dirty*" provides both *positive* and *negative* sentiment scores for the word "*extremely*" and averaging these values result in the first sentence having a lower overall *negative* polarity than the second sentence. Sentiment scores also badly influenced the classification performance like in the tweet "*@obeseziall or say that someone should get raped. That's just sick and twisted and person has to be really mean and sadistic to say that*", we obtain sentiment scores as −0.07, −0.146, 0.115, 0.032, 0.102, 0.156, 0.177, 0.0, 0.063, 0.25, 1.25, indicating overall high *negative* polarity but it actually belongs to Class 1 (*neither offensive nor hate speech*). In Module 5, a drawback of calculating sentence vectors in this way is that we lose the relative ordering of the words in the sentence. Also, the *Word2Vector* training was done for 100 epochs. More number of iterations could lead to better performance as the dataset size grows.

Training data with equal distribution from all three classes tend to outperform the other cases. Training with the best individually performing datasets (D2 and D5) also has similar results. These two give the highest accuracy during testing (as shown in Fig. 6).

6 Conclusion and Future Work

Here, we have experimented with a text classification system for classifying hate speeches in twitter using a neural network model. The classifier assigns to each

Table 2 Confusion matrices for all case studies

	Case 1 (Accuracy 76.57)				Case 2 (Accuracy 69.67)				Case 3 (Accuracy 71.05)			
Dataset	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total
Neutral	4043	222	492	**4762**	631	45	24	**700**	656	20	24	**700**
Offensive	231	1349	930	**2560**	88	353	259	**700**	84	357	259	**700**
Hate	51	160	1689	**1900**	59	162	479	**700**	62	159	479	**700**
Total	**4380**	**1731**	**3111**	**9222**	**778**	**560**	**762**	**2100**	**802**	**536**	**762**	**2100**
	Case 4 (Accuracy 70.57)				Case 5 (Accuracy 76.62)				Case 6 (Accuracy 73.99)			
Dataset	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total
Neutral	670	25	5	**700**	3388	120	34	**3542**	636	42	19	**697**
Offensive	93	441	166	**700**	343	1583	459	**2385**	94	239	43	**376**
Hate	76	253	371	**700**	138	596	566	**1300**	62	76	81	**219**
Total	**839**	**719**	**542**	**2100**	**3869**	**2299**	**1059**	**7227**	**792**	**357**	**143**	**1292**

Table 3 Confusion matrices for all modules

	Module 1				Module 2				Module 3			
Dataset	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total
Neutral	4231	393	105	**4729**	3877	578	274	**4729**	4172	374	153	**4699**
Offensive	529	1966	706	**3201**	370	1696	1135	**3201**	383	1683	1103	**3169**
Hate	223	791	744	**1758**	136	395	1227	**1758**	140	386	1225	**1751**
Total	**4983**	**3150**	**1555**	**9688**	**4383**	**2669**	**2636**	**9688**	**4695**	**2443**	**2481**	**9619**

	Module 4				Module 5			
Dataset	Neutral	Offensive	Hate	Total	Neutral	Offensive	Hate	Total
Neutral	3558	1053	1209	**5820**	3619	538	570	**4727**
Offensive	1396	834	1653	**3883**	967	611	1613	**3191**
Hate	627	364	932	**1923**	457	296	1001	**1754**
Total	**5581**	**2251**	**3794**	**11,626**	**5043**	**1445**	**3187**	**9672**

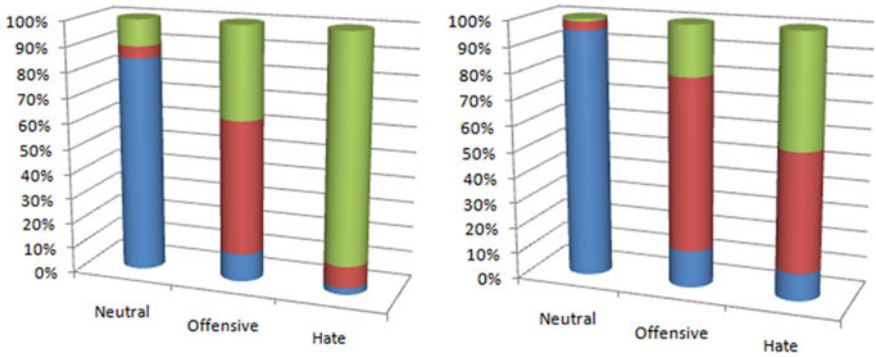


Fig. 6 Overall accuracies (in %) for *Case 1* and *Case 5* of module 3

tweet one out of three different classes: *neutral, offensive,* and *hate*. Working with a dataset of over 12k tweets, we experimented with different features including *n-gram, sentiment and word2Vec*. Our best model of using the *4-gram* and *sentiment* features performed with an overall accuracy of 73.6%. The experiments performed here were using a convolutional neural network (CNN), which leaves several other deep neural network architectures to be tried.

Acknowledgements The work reported here is supported by the grant in aid from the IMPRINT-II sponsored research project, "Sevak—An Intelligent Indian Language Chatbot", funded by SERB, Government of India.

References

1. Waseem Z, Hovy D (2016) Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop
2. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media, ACL, pp 1–10
3. Badjatiya P, Gupta S, Gupta M, Varma V (2017) Deep learning for hate speech detection in tweets. In: 26th WWW conference, pp 759–760
4. Bohra A, Vijay D, Singh V, Akhtar SS, Shrivastava M (2018) A dataset of Hindi-English code-mixed social media text for hate speech detection. In: Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media, pp 36–41
5. Gamback B, Sikdar UK (2017) Using convolutional neural networks to classify hate-speech. In: Proceedings of the first workshop on abusive language online, p 8590
6. Xiang G, Fan B, Wang L, Hong J, Rose C (2012) Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: Proceedings of the 21st ACM international conference on Information and knowledge management, Maui, HI, USA. ACM, pp 1980–1984
7. Davidson T, Warmsley D, Macy M, Weber I (2017) Automated hate speech detection and the problem of offensive language. In: Proceedings of ICWSM

8. Kwok I, Wang Y (2013) Locate the hate: detecting tweets against blacks. In: des Jardins M, Littman ML (eds) AAAI, Bellevue, Washington. AAAI Press, pp 1621–1622
9. Gitari ND, Zuping Z, Damien H, Long J (2015) A lexicon-based approach for hate speech detection. Int J Multimedia Ubiquitous Eng 10(4):215–230

Impact of Visual Representation of Audio Signals for Indian Language Identification

Himanish Shekhar Das and Pinki Roy

Abstract In multilingual-based Indian society, the capability to identify and classify a spoken language is an imperative task. Recently, deep learning algorithms, such as convolutional neural networks, seem to be effective in achieving better accuracy for images. Variants of convolutional neural networks such as VGG network, which are trained on large-scale datasets have also demonstrated to be excellent for the task of image recognition. Using a pretrained convolutional neural network, which acts as feature extractor, provides an alternative to the handcrafted features. These features involve the effective learning transformation of raw speech signals to a new representation which can be effectively used in various learning mechanisms. In this paper, initially speech signal is converted into a mel-spectrogram image which includes the multi-frequency band signals with respect to time. Then, the feature extraction process is carried out using VGGNet-based feature extractor to extract deep linguistic features from the spectrogram images. Finally, softmax regression, random forest, and support vector machine algorithms are being utilized for language identification to calculate language score for each language. The simulation is conducted on a record database which consists of four Indian languages with special emphasis on Northeastern languages. The languages are Assamese, Bengali, Hindi, and Indian English, respectively. The simulation results show that the support vector machine model achieved supreme accuracy up to 96.50% as compared to other models.

Keywords Language identification · Spectrogram images · Machine learning algorithms · Convolutional neural networks

1 Introduction

Spoken language identification (LID) is characterized as the way toward perceiving the language from a given discourse articulation, supported for multilingual speech processing applications [1]. Because of its various applications, LID has received

H. S. Das (✉) · P. Roy
Department of Computer Science and Engineering, National Institute of Technology Silchar, Silchar, Assam 788010, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_32

much attention to major research areas such as computational sciences, namely speech recognition, speaker verification, machine translation, and retrieval of spoken document [2]. In a multilingual country like India, the automatic recognition of speech by a machine is a challenging task as most of the people in India are habituated in using common multilingual words while speaking [3–5]. In multilingual countries, the LID task becomes more challenging due to the mutual phoneme set of languages [6]. Moreover, in an LID model, the languages either can be preclassified into various sublanguages for classifying the huge number of target languages in multilingual countries [7].

In audio signal processing field, the traditional LID systems utilize the domain specific expert knowledge for extracting the handcrafted features from speech samples [8]. For feature extraction process, majority of the LID scheme uses perceptual linear prediction (PLP), linear prediction cepstral coefficient (LPCC), and mel-frequency cepstral coefficient (MFCC), etc., while algorithms like hidden Markov model (HMM), Gaussian mixture model (GMM), support vector machine (SVM), artificial neural network (ANN), etc., are being used as the back-end classifier. Due to the robustness in such task, most of the time MFCC is generally utilized as an acoustic feature for feature extraction [9, 10]. Deep learning techniques have shown promising results and overshadowed GMM and HMM era [11]. Since the evaluation of optimal hand crafted features to attain a specific goal is a bit tough, in current time, many different efforts have been made to learn the effective features using deep neural network (DNN).

For the speech signals, one of the important information sources is spectrogram, which is the visual representation of audio signals. For speech analysis, spectrogram is widely utilized which represents the contextual variations in a good manner [12]. To get a better idea about the applicability of deep learning methods in the area of speech signal processing, to perform speech recognition, many researchers have done lots of experimental work applying the deep learning techniques. Features extracted using the raw input from convolutional neural network (CNN) have been used in various fields of research, viz. speech [13], emotion [14], and gesture recognition [15] very well. CNN is effective because it has the ability to extract the frequency shift and to exploit the frequency, time correlation in human speech signal [16, 17]. Compared to other variants of CNN, VGGNet is gaining more attention that has several advantages. The architecture of VGGNet is very simple and the architecture is very good for benchmarking on a particular task [18]. Like VGGNet, there are other pretrained networks are available freely on the Internet, so it is commonly used for various applications [19].

In the proposed approach, inputs are spectrogram images which have been widely used in the audio classification tasks. CNNs are very good at abstracting data which is used in the area of image recognition as well as speech recognition. In this paper, to analyze the impact of image features from spectrograms, we have taken VGGNet as one of the variants of CNN, which can efficiently exploit invariance presented in spectrogram using their pooling and convolutional operations. To verify the efficiency of deep features from images, a comparison for the standard handcrafted base-line

feature such as using statistics of the mel-spectrograms with the same set of classifiers have also been performed. The major contributions of this paper are as follows:

- Development of CNN-based softmax regression (SR), random forest (RF), and SVM for LID system.
- Analyzing the benefits of VGGNet for feature extraction to get high sensitive feature level representation.
- Improvement of language identification model using machine learning algorithms to calculate language score for each language.

The organization of this paper is as follows: Sect. 2 analyzes the related work. The proposed methodology is covered in Sect. 3. Performance analysis is explained in the results and discussion which is in Sect. 4. The conclusion of the work is included in Sect. 5.

2 Related Works

Many techniques have been proposed for automatic LID in this literature. For spoken LID, Deshwal et al. [20] have used different hybrid robust feature extraction techniques in their work. Various combinations of acoustic features like MFCC, PLP, and SDC have been tested for LID approach using feedforward deep neural network. An end-to-end short duration utterances-based LID model was presented by Ma et al. [21]. This method was designed with LSTM neural networks, which was implemented in the speech LID application. A phase vocoder-based time scale alteration scheme was used to enhance the SLD accuracy with short utterances for diminishing the test utterance speech rate. A supervised learning method was designed by Mandal et al. [22] for LID. At low resources word level Bengali, English code-mixed data given from social media. They utilized two schemes of word encoding, specifically root phone based and character based to train the deep LSTM module. In our previous work also, we have introduced the efficiency of excitation features, prosodic features [23], as well as a deep bottleneck feature [24] using various deep learning architectures. For automatic spoken LID, Montalvo et al. [25] proposed one front-end technique which was based on the spectrogram of speech signals. To capture the spectrogram content, as a texture descriptor, the local phase quantization was utilized. The LID for audio spectrograms utilizes spectrograms of raw audio signals by Revay and Teschke [9]. Proposed model uses spectrograms as input to a CNN. Language identification for audio spectrograms (LIFAS) shows great improvement in terms of accuracy for binary language classification as well as multiclass classification, respectively. The spectrogram-based textual descriptors and evolutionary feature selection were studied by Chowdhury et al. [26] for Indian language identification. At first, the speech signal is transformed into the spectrogram image representation. Features extracted from spectrogram images using local binary pattern histogram Fourier, completed local binary pattern as well as from discrete

wavelet transform. To construct the classification model, grey wolf optimization feature selection algorithm is utilized.

3 Proposed Methodology

In this paper, a deep feature-based LID is developed for recognizing the languages from the unknown speech signal. In earlier LID systems, the process of feature selection was manual. Also, the process of feature extraction was algorithm specific. In contrast, in deep learning networks, the process of learning high level features from data is an automatic process. The downside is, training these networks is a costly affair since these may contain a wide range of parameters, which in turn requires massive amounts of data, and to gather such massive amount of data, one requires a huge number of speech samples. The need for data in large quantity and computational resources in abundance is not practical in certain scenarios. One mitigation policy is to adapt pretrained networks, which follow transfer learning. The aim of transfer learning is to execute knowledge gained from one problem to another problem which is related. The benefit is that the new problem does not have to learn weights from the scratch since the weights corresponding to features are already learned by the pretrained network. Here, the output of the VGG network, which is just before the last fully connected layer, composes the set of features which are of interest. In this work, initially each audio file is taken from the database and then it is transformed into spectrogram images. Then those images are fed into CNN. The features are extracted by using VGGNet architecture. Features from CNN are further feed into three algorithms, i.e., SR, RF, and SVM for classification for identifying the languages. In this work, each model has been trained with 4000 images and tested with 1000 images. Figure 1 shows the workflow of the proposed approach. Along with that, a comparison has been performed for base-line handcrafted features like mel-frequency cepstrum coefficients (MFCC) against deep features extracted from mel-spectrogram images, for the same set of classifiers. To conduct the experimental analysis for handcrafted feature, 13 MFCC features are derived from a speech utterance.

3.1 Mel-Spectrogram Representation

The visual representation of frequency spectrum in a sound or other signal concerning time is known as a spectrogram. It is signified by a two-dimensional graph, the vertical axis is denoted by frequency and the horizontal axis is denoted by time. Figure 2 shows the sample spectrogram images. Using a Fourier transform, the frequency spectrum of a signal can be created from a time series signal. To calculate the magnitude of frequency spectrum, the fast Fourier transform can be used for a fixed moment in time. Acoustic time–frequency representation of a sound is vividly expressed by a log-mel spectrogram. The process of computing log-mel spectrogram

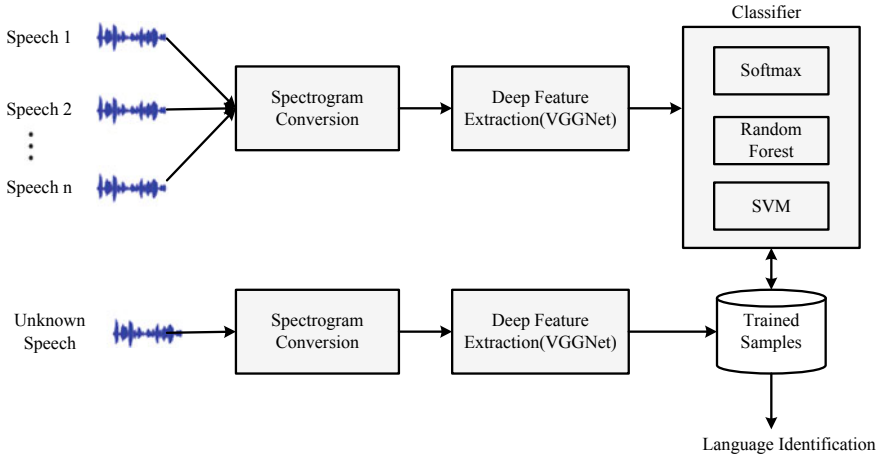


Fig. 1 Proposed model for language identification

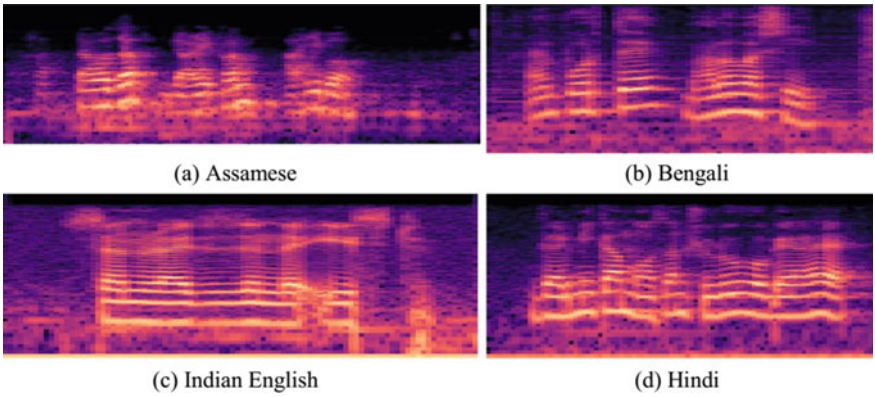


Fig. 2 Sample mel-spectrogram images

is a two-step process. First, the fast Fourier transform is computed over preprocessed audio signal as represented by the following equation:

$$X_i(k) = \sum_{n=0}^{N-1} x_i(n)h(n)\mathrm{e}^{-j\frac{2\pi}{N}n} \quad k = 0, \ldots, N-1 \tag{1}$$

Then, the mapping of its spectral amplitude to the mel-scale of the perceptual excitation is done with the filter bank, and the mel-filter bank transforms the spectrum to the mel-spectrum. Human hearing frequencies are perceived to be of mel-scale. Thus, the mel-scale is used for measuring the tone of a subjective frequency or pitch. The mel-frequency scale can be described by the following equation:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

where f is the frequency in hertz and m is the number of mels.

3.2 Deep Feature Extraction Using CNN

In our previous work, we have tested various handcrafted features as well as a combination of bottleneck features in deep learning environments. In this work, to upgrade feature level from handcrafted feature to deep feature, we have utilized convolutional neural network. CNNs are extensively used for image classification and detection. Images of pixels are 2D matrix on which we run CNN to either classify the image or recognize the image. There is a sequence of layers in CNN and each layer converts one volume to another through differentiable function.

In this paper, input layer holds the raw image with depth 3, height 224, and width 224. The architecture of CNN model is given in Fig. 3. The convolution layer registers the output volume by figuring the dot product between image patch and all filters. Using transfer learning for feature extraction from VGG-16 architecture, $7 \times 7 \times 512$ number of features have been extracted. To extract deep features of the VGG16 network, we in essence "chop off" the network just before the fully connected (FC) layers. The final layer in the network which is a max pooling layer, whose output is of the shape of $7 \times 7 \times 512$ implying that there are 512 filters. Therefore, these $7 \times 7 \times 512 = 25{,}088$ values can be treated as a feature vector which quantifies the contents of spectrogram image.

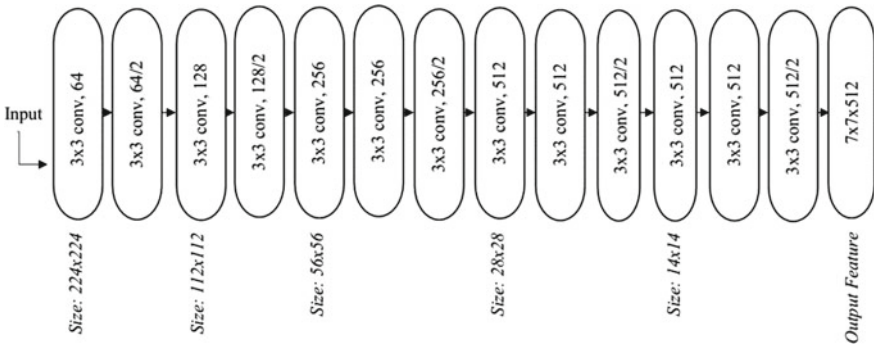


Fig. 3 Architecture of VGG-16 network

3.3 List of Comparative Algorithms

Softmax Regression

Extracted feature from CNN is adopted as input to the softmax classifier in the proposed method. In classification problems, softmax or multinomial logistic regression [27] comes into picture when the possible values of class label are more than two. It is a generalization of logistic regression to those cases where we have to deal with multiple classes. The binary logistic regression is the generalization of this softmax classifier to multiple classes. A probabilistic interpretation and more intuitive output are given by the softmax classifier. For any test input say x, we can calculate the probability $P(l = z|x)$ for each value of $z = \{1, 2, \ldots, n_l\}$, where n_l is the number of labels. A vector of Y real numbers is taken as the input in the function and the proportionate exponentials of the input numbers are being normalized into a probability distribution which is consisting of Y probabilities.

Random Forest

This classifier uses supervised learning approach, which can be applied in solving the problems, including classification as well as regression. It is a collection of many decision trees where each decision tree is trained on a different set of the observations. Splitting of the nodes in each tree is based on considering a limited number of features. The final prediction of the random forest is done by taking the average of all the predictions from each individual tree. The feature extracted from VGGNet act as input to the random forest classifier in the proposed method. For any given speech characteristic variable x, the optimal classification result is determined by the vote of each decision tree classifier. The steps involved in generating the RF are listed as follows:

- Step 1: If the number of decision tree is k then using bootstrapping, out of them only k number of subset samples can be taken out randomly without replacement. Each of these subsets will be acting as the training set to grow a tree. Thereafter the sample which is not selected earlier is used perform the testing of the trained model.
- Step 2: Assume n features are chosen randomly out of say N features. To get the best classification from this approach, selected n features are chopped off in different groups using optimized split approach. The value of n does not change during the whole process and the process is repeated till the whole tree has been created completely.
- Step 3: Each single tree can grow up to a maximum without performing any cutting.

The generated trees constitute the RF, and in this paper, the RF algorithm is used to classify the speech data. The results of each decision tree determine the output of the RF. One of the major advantages of this approach is that it does not suffer from an overheating problem, since it takes the average of all the predictions it gets from each decision tree.

Support Vector Machine

In multilingual language identification, binary classification algorithms can be turned into multiclass classification algorithms. SVM can be used not only for classification problems, it has also performed well for regression problems. There are two methods for multiclass classification of SVM, one is one against one approach (ONO) and the other is one against all approach (ONA). ONA approach is preferred when we want to convert N classes learning problem into many two class learning problems. Using a hyperplane with the largest amount of margin, the classifier separates data points and finds an optimal hyperplane which has the capability to classify new samples. In the classification process, the kernel has a major role. The input samples are set to the n-dimensional feature space by the kernel. The kernel plays a major role in the process of classification, and it sets the input samples to *the n-dimensional feature space*. Role of kernel is to take the data from a low-dimensional input space and transform that into a higher-dimensional space. For this work, radial basis function (RBF) kernel has been used, though there are some disadvantages of SVM which include high training time for large datasets and sensitivity toward the type of kernel used.

4 Results and Discussion

The proposed method for Indian language identification with special emphasis on Northeastern languages [23, 24] is evaluated with recorded database in this paper. The motivation behind working on the selected four languages is that these languages are mostly spoken in Assam and our main focus is to work with Northeastern languages of India. The dataset is at sentence level where for each language a particular sentence has been recorded. Table 1 represents the full description of the speech database.

Table 1 Database description

Language	No of speakers	No. of times sentence rehashed by each speaker	Total no. of training samples	Total no. of testing samples	Sampling frequency	Resolution
Assamese	50	25	1000	250	16 KHz	16 bps
Bengali	50	25	1000	250		
English	50	25	1000	250		
Hindi	50	25	1000	250		
Total	200		4000	1000		

4.1 Evaluation Metrics

For the classification task of the cascading CNN + softmax, CNN + RF, CNN + SVM, following metrics we have used to check the performance of each classifier. From the confusion matrix, we get the following parameters which are true positive (TP), true negative (TN), false negative (FN), and false positive (FP), respectively. For the evaluation of this work, various performance measures considered are given in the following equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{3}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$F\text{-score} = \frac{(1 + \alpha^2)\text{Recall} * \text{Precision}}{\alpha^2(\text{Recall} + \text{Precision})} \tag{6}$$

Equal error rate (EER) is defined as the point in which the false acceptance rate (FAR) is equal to the false rejection rate (FRR). It is a measure used to calculate the accuracy of language identification task.

$$\text{FAR} = \frac{\text{Number of false acceptance}}{\text{Number of identification attempt}} \sim \frac{\text{FP}}{(\text{FN} + \text{FP})} \tag{7}$$

$$\text{FRR} = \frac{\text{Number of false rejection}}{\text{Number of identification attempt}} \sim \frac{\text{FN}}{(\text{FN} + \text{FP})} \tag{8}$$

Tables 2, 3, and 4 show the confusion matrices for CNN + softmax, MFCC + softmax, CNN + RF, MFCC + RF, CNN + SVM and MFCC + SVM, respectively. Tables 5, 6, 7, and 8 give the performance measure with regard to accuracy, precision, recall, and F-measure, for the proposed approach as well as for the comparison with four Indian languages.

Table 2 Confusion matrices for CNN + softmax and MFCC + softmax

Languages	Deep features				MFCC features			
	Assamese	Bengali	English	Hindi	Assamese	Bengali	English	Hindi
Assamese	228	10	4	8	184	27	16	23
Bengali	4	240	3	3	25	186	21	18
English	5	4	237	4	23	21	185	21
Hindi	7	5	4	234	25	20	21	184

Table 3 Confusion matrices for CNN + RF and MFCC + RF

Languages	Deep features				MFCC features			
	Assamese	Bengali	English	Hindi	Assamese	Bengali	English	Hindi
Assamese	231	9	2	8	188	26	14	22
Bengali	4	241	2	3	22	191	20	17
English	4	5	236	5	22	18	192	18
Hindi	3	4	3	240	23	18	18	191

Table 4 Confusion matrices for CNN + SVM and MFCC + SVM

Languages	Deep features				MFCC features			
	Assamese	Bengali	English	Hindi	Assamese	Bengali	English	Hindi
Assamese	233	8	2	7	199	22	11	18
Bengali	3	244	1	2	17	204	16	13
English	2	3	243	2	18	14	205	13
Hindi	2	2	1	245	19	14	13	204

Table 5 Accuracy performance (in %)

Languages	CNN + Softmax	CNN + RF	CNN + SVM	MFCC + Softmax	MFCC + RF	MFCC + SVM
Assamese	96.2	97.0	97.6	86.1	87.1	89.5
Bengali	97.1	97.3	98.1	86.8	87.9	90.4
English	97.6	97.9	98.9	87.7	89	91.5
Hindi	96.9	97.4	98.4	87.2	88.4	91.0

Table 6 Precision performance (in %)

Languages	CNN + Softmax	CNN + RF	CNN + SVM	MFCC + Softmax	MFCC + RF	MFCC + SVM
Assamese	93.4	95.4	97.0	71.6	73.7	78.7
Bengali	92.6	93.0	94.9	73.2	75.5	80.3
English	95.5	97.1	98.3	76.1	78.9	83.7
Hindi	93.9	93.7	98.0	74.8	77	82.3

Table 7 Recall performance (in %)

Languages	CNN + Softmax	CNN + RF	CNN + SVM	MFCC + Softmax	MFCC + RF	MFCC + SVM
Assamese	91.2	92.4	93.2	73.6	75.2	79.6
Bengali	96.0	96.4	97.6	74.4	76.4	81.6
English	94.8	94.4	97.2	74	76.8	82
Hindi	93.6	96.0	98.0	73.6	76.4	81.6

Table 8 *F*-measure performance (in %)

Languages	CNN + Softmax	CNN + RF	CNN + SVM	MFCC + Softmax	MFCC + RF	MFCC + SVM
Assamese	92.3	93.9	95.1	72.6	74.5	79.1
Bengali	94.3	94.7	96.2	73.8	76.0	81.0
English	95.1	95.7	97.7	75.1	77.7	82.8
Hindi	93.7	94.8	96.8	74.2	76.7	82.0

Experimental results from Tables 5, 6, 7, and 8, inference that SVM achieves better results with deep features extracted from spectrogram images, as well as deep features from mel-spectrogram images perform better compared to handcrafted feature like MFCC. Tables 5, 6, 7, and 8 also signify in comparison with the accuracy obtained for Bengali and Assamese language, English and Hindi language shows better recognition ratio. Tables 2, 3, and 4 reflect that CNN + softmax, CNN + RF, and CNN + SVM achieve 93.90%, 94.80%, and 96.50% accuracy, respectively. From Fig. 4 also we can conclude that SVM shows promising results with deep features, and it also reflects the improvement when compared with MFCC features. Based on the results obtained from the proposed LID models, it shows that for the recognition of Indian languages, deep feature-based SVM model outperforms other state-of-the-art methods.

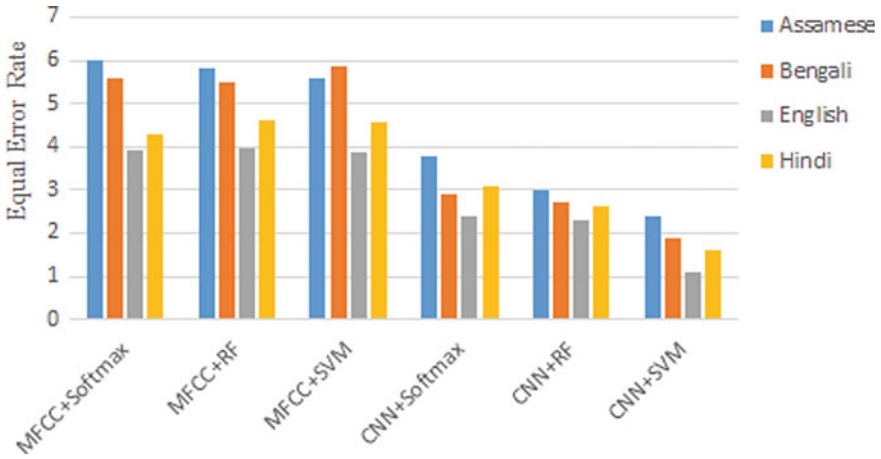


Fig. 4 Equal error rate performance comparison

5 Conclusion

In this paper, three types of cascading model have been proposed for Indian languages especially for Northeastern languages. To verify the usefulness of image-based feature for Indian language identification, deep feature-based various hybrid models have been proposed which are CNN-based softmax, random forest, and support vector machine. Initially, speech samples from the speech dataset have been converted into spectrogram images and those images act as input for VGG-16 network to extract deep features. The features extracted from the CNN model have been further fed into the various algorithms for precise language identification. The performance measures such as accuracy, precision, recall, F-measure, and error rate were calculated for softmax, random forest, and SVM. Experimental results show that CNN + softmax, CNN + RF, and CNN + SVM achieve 93.90%, 94.80%, and 96.50% accuracy, respectively, from which we can conclude that SVM performs better with deep features from spectrogram images, with high accuracy and less error rate.

References

1. Muthusamy YK, Barnard E, Cole RA (1994) Automatic language identification: a review/tutorial. IEEE Signal Process Mag 11(4):33–41
2. Bekker AJ, Opher I, Lapidot I, Goldberger J (2016) Intra-cluster training strategy for deep learning with applications to language identification. In: 26th international workshop on machine learning for signal processing (MLSP). IEEE, pp 1–6
3. Lounnas K, Abbas M, Teffahi H, Lichouri M (2019) A language identification system based on voxforge speech corpus. In: International conference on advanced machine learning technologies and applications. Springer, Cham, pp 529–534
4. van der Lee C, van den Bosch A (2017) Exploring lexical and syntactic features for language variety identification. In: Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial), pp 190–199
5. Geng W, Wang W, Zhao Y, Cai X, Xu B, Xinyuan C (2016) End-to-end language identification using attention-based recurrent neural networks. In: Interspeech, pp 2944–2948
6. Watanabe S, Hori T, Hershey JR (2017) Language independent end-to-end architecture for joint language identification and speech recognition. In: IEEE automatic speech recognition and understanding workshop (ASRU). IEEE, pp 265–271
7. Thenmozhi D, Kayalvizhi S, Aravindan C (2018) A machine learning approach to Indian native language identification. In: FIRE (working notes), pp 68–76
8. Aarti B, Kopparapu SK (2018) Spoken Indian language identification: a review of features and databases. Sādhanā 43(4):53
9. Revay S, Teschke M (2019) Multiclass language identification using deep learning on spectral images of audio signals. arXiv preprint arXiv:1905.04348
10. Deshwal D, Sangwan P, Kumar D (2019) Feature extraction methods in language identification: a survey. Wireless Pers Commun 107(4):2071–2103
11. Das HS, Roy P (2019) A deep dive into deep learning techniques for solving spoken language identification problems. In: Intelligent speech signal processing. Academic Press, Cambridge, pp 81–100
12. Mukherjee H, Dhar A, Obaidullah SM, Phadikar S, Roy K (2020) Image-based features for speech signal classification. Multimedia Tools Appl 1–17

13. Tatulli E, Hueber T (2017) Feature extraction using multimodal convolutional neural networks for visual speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2971–2975
14. Gupta S, De K, Dinesh DA, Thenkanidiyoor V (2019) Emotion recognition from varying length patterns of speech using CNN-based segment-level pyramid match kernel based SVMs. In: National conference on communications (NCC). IEEE, pp 1–6
15. Ahlawat S, Batra V, Banerjee S, Saha J, Garg AK (2019) Hand gesture recognition using convolutional neural network. In: International conference on innovative computing and communications. Springer, Singapore, pp 179–186
16. Shukla S, Mittal G (2019) Spoken language identification using ConvNets. In: European conference on ambient intelligence. Springer, Cham, pp 252–265
17. de Benito-Gorron D, Lozano-Diez A, Toledano DT, Gonzalez-Rodriguez J (2019) Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. EURASIP J Audio Speech Music Process 2019(1):9
18. Rezende E, Ruppert G, Carvalho T, Theophilo A, Ramos F, de Geus P (2018) Malicious software classification using VGG16 deep neural network's bottleneck features. In: Information technology-new generations. Springer, Cham, pp 51–59
19. Wu Z, Shen C, Van Den Hengel A (2019) Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recogn 90:119–133
20. Deshwal D, Sangwan P, Kumar D (2020) A language identification system using hybrid features and back-propagation neural network. Appl Acoust 164:107289
21. Ma Z, Yu H, Chen W, Guo J (2018) Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features. IEEE Trans Veh Technol 68(1):121–128
22. Mandal S, Das SD, Das D (2018) Language identification of Bengali-English code-mixed data using character & phonetic based LSTM models. arXiv preprint arXiv:1803.03859
23. Das HS, Roy P (2019) Optimal prosodic feature extraction and classification in parametric excitation source information for Indian language identification using neural network based Q-learning algorithm. Int J Speech Technol 22(1):67–77
24. Das HS, Roy P (2020) Bottleneck feature-based hybrid deep autoencoder approach for Indian language identification. Arab J Sci Eng 45:3425–3436
25. Montalvo A, Costa YM, Calvo JR (2015) Language identification using spectrogram texture. In: Iberoamerican Congress on pattern recognition. Springer, Cham, pp 543–550
26. Chowdhury AA, Borkar VS, Birajdar GK (2020) Indian language identification using time-frequency image textural descriptors and GWO-based feature selection. J Exp Theor Artif Intell 32(1):111–132
27. Yin M, Zeng D, Gao J, Wu Z, Xie S (2018) Robust multinomial logistic regression based on RPCA. IEEE J Sel Top Signal Process 12(6):1144–1154

Bi-Lingual TDNN-LSTM Acoustic Modeling for Limited Resource Hindi and Marathi Language ASR

Ankit Kumar and Rajesh Kumar Aggarwal

Abstract It is a need of time to build an automatic speech recognition (ASR) system for low-resource languages. India is a multilingual country with more than 5000 languages, and out of which 22 languages are official. The ASR systems for Indian languages are still in the infancy stage, mainly due to resource deficiency (e.g., lack of transcribed speech data, pronunciation lexicon, and text data). Deep neural networks (DNNs) have significantly improved the performance of the ASR system of low-resource languages. The merging data of various languages is a commonly used trend to train the multilingual DNN acoustic model. In multilingual training, hidden layers act as a global feature extractor. The multilingual ASR systems work better for similar kinds of sources and target languages. In this work, we use two low-resource Indian languages, namely Hindi and Marathi. Both languages are closely related and belong to the same Indo-Aryan family. Both languages include various common phonemes and words. Various state-of-the-art experiments were performed using different acoustic modeling and language modeling techniques. Experiments demonstrated that multilingual ASR systems consistently outperform monolingual ASR systems for both Hindi and Marathi languages.

Keywords Automatic speech recognition (ASR) · Multilingual ASR · TDNN · LSTM · Acoustic modeling

1 Introduction

In the last few years, ASR has been improved tremendously, which is used commercially for resource-rich languages like English. The reason behind the success of ASR systems like Siri, Alexa, and Google Voice is the availability of a huge amount of training data [1]. Thousands of hours of transcribed speech data have been used to

A. Kumar · R. K. Aggarwal
Department of Computer Engineering, National Institute of Technology, Kurukshetra, India

A. Kumar (✉)
School of Computing Science & Engineering, Galgotias University, Greater Noida, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_33

build the acoustic model of these ASR systems [15, 42]. DNN-based acoustic modeling is a current trend in large vocabulary continuous speech recognition (LVCSR) systems as it estimates millions of parameters without overfitting [1]. For English and other European languages, large speech and language resources are easily available to apply deep learning acoustic modeling techniques [3, 6, 14, 16, 28]. On the other hand, there is a long list of languages which qualified to be low-resource languages due to the lack of transcribed speech corpus, availability of text resources, and pronunciation dictionary [5]. The quick development of the ASR system for low-resource language is the major hurdle of the ASR community [19]. India is a multilingual nation with more than 500 languages, out of which 22 languages are official [9]. The major Indian languages are low-resource languages, which attracts the attention of various industries and academicians. However, till now, only a few attempts have been recorded for low resource Indian languages [10, 18, 23–26, 29, 49]. People from the rural part of India will become part of the information revolution by using ASR applications in their native language.

The robustness of any ASR system highly depends upon the amount of available training data [1]. The collection of data with their proper transcription is not economical and consumes an immense amount of time. To learn a new language, human borrows the information from the other languages due to the overlapping phonetic system of language as all humans share the same vocal tract architecture and phonetic system of languages [19]. This process is known as transfer learning. The same concept has been used in this work, in which multilingual DNN training has been done for Marathi and Hindi language with separate output layer for each language. This method is generally known as block-softmax [34]. Recently, multilingual ASR systems with deep learning-based acoustic modeling for low-resource languages have become more popular as they have given consistent better performance [13, 17, 45, 48]. The multilingual training significantly reduces the word error rate (WER) 3–5% compared to monolingual deep learning-based acoustic models [47]. In such architecture, multilingual speech data is available in an interleaved fashion in hidden layers in congruence with language-specific output layers. There are only a few attempts have been recorded for the multilingual ASR system of Indian languages [5, 12, 37, 43, 49]. All works in [5, 12, 37, 43, 49] are done for Tamil, Telugu, and Gujarati language.

In this research paper, we grouped Hindi (2.5 h) and Marathi (5.5 h) data together to train the multilingual system on deep learning acoustic models, in which hidden layers work as a global feature extractor. Therefore, hidden layers share information from other languages in order to benefit each other [17, 47]. In all experiments, we considered two low-resource languages Hindi and Marathi, for various acoustic modeling. Hindi and Marathi both are Indo-Aryan languages which are closely related and share many common phone set. The combined Hindi and Marathi dataset will lead to better apply the transfer learning by using the small dataset to improve the performance. In all experiments, we observed consistent improvement compared to the mono-lingual-based acoustic modeling.

The remaining part of the paper is organized as follows: Sect. 2 briefly describes the acoustic models, and Sect. 3 explains the architecture of the multilingual ASR

system. Section 4 describes the speech corpus. Section 5 gives the experimental setup details. Section 6 covers the experimental part of the paper, and Sect. 7 is the conclusion of the proposed system.

2 Acoustic Models Exploiting Variable-Length Contextual Information

DNNs only take a fixed-length context window of the frame as input and fail to exploit long-range correlations in speech signal [50]. Generally, DNN uses 9–13 frames window (left/right context of 4–6 frames) of features as an input to exploit the contextual information from neighboring frames to enhance the system accuracy. These fixed-length context windows are not good enough to exploit contextual information as optimal window size may vary for different speaking speed and phones [50]. On the other side, time-delay neural network (TDNN) and long short-term memory (LSTM) models are more able to effectively exploit the variable length of contextual information. The detailed discussions are as follows:

2.1 Sub-sampled TDNN

TDNN model can effectively exploit the variable-length contextual information. TDNN was firstly reported in ASR for isolated word recognition task in 1990 [27]. In contrast to convolutional neural network (CNN), TDNN convolves along with both frequency and time axis, resulting in a better ability to handle temporal context [50]. In the TDNN network, the input layer learns the initial transform in a narrow context, and higher layers perform activation from a wider temporal context [22, 31]. Every hidden layer in TDNN architecture operates on a different temporal resolution, and this resolution subsequently increases layer by layer. Due to this significant feature, the deeper layer of the TDNN network can learn a wider temporal relationship. Typically, hidden activations are computed in TDNN at all time steps by sampling together contiguous frames at each hidden layer. Thus, the computational load of such a network becomes very high. Due to the processing of contiguous frames, there are large overlaps between input contexts. The neighboring activations become highly correlated as well. They can be sub-sampled. The computational load in TDNN can be significantly reduced by using sub-sampling TDNN layers output [30, 31]. TDNN network works better if splicing of wide frames at deeper layers of the network [39]. Due to the selective activation function of time steps, we reduce the overall computational load in forwarding pass and back-propagation (Fig. 1).

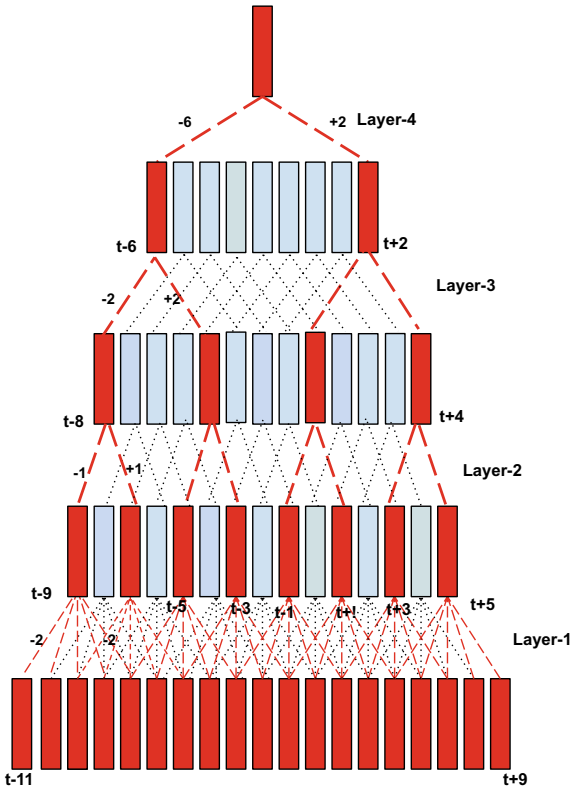


Fig. 1 TDNN Architecture based on sub-sampling. The notation $\{-6, +2\}$ denotes splicing of t-6 frame with t+2 frame to apply activation function. The frame t-2 to t+2 $\{-2, -1, 0, +1, +2\}$ or more compactly as [-2,+2] is spliced together at input layer. After that splicing of frame offset at three hidden layers are $\{-1, +1\}$, $\{-2, +2\}$ and $\{-6, +2\}$

2.2 Long Short-Term Memory (LSTM) Model

On the contrary to DNN, recurrent neural networks (RNNs) can encode the sequence history in their internal state, which gave a potential to predict the next phoneme by using previously observed speech features. This basic RNN architecture suffered from a gradient issue, which was later on resolved by introducing LSTM (Fig. 2).

The LSTM contains memory cells to store the temporal state of the network in the recurrent hidden layer. Except for the memory cell, each block of LSTM contained an input gate, output gate, and forget gate to control the flow of information in the network [40]. By using these gates, the gradient can be propagated in the network in a more stable fashion over a long span of time. The flow of input activation into a memory cell is controlled by an input gate, and the output flow of the memory cell is controlled by an output gate [40]. The forget gate resets the cell memory every time.

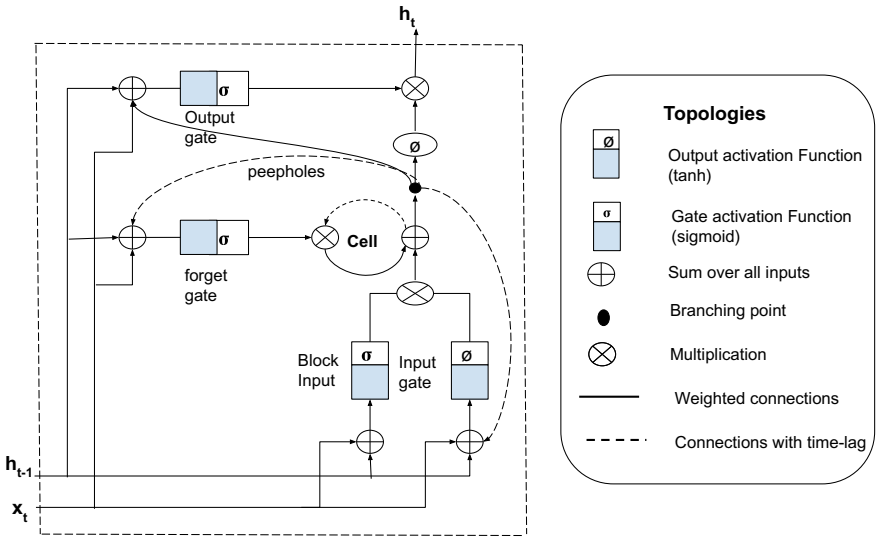


Fig. 2 LSTM memory block (cell) representation used in hidden layers of RNN

The peepholes connections are added to learn the precise timing. Let x^t be the input vector at time stamp t, then computation at each block of LSTM can be described as:

$$a^t = \phi(\widehat{a}^t) = \phi(W_c X^t + R_c h^{t-1} + b_c) \tag{1}$$

$$i^t = \sigma(\widehat{i}^t) = \sigma(W_i X^t + R_i h^{t-1} + p_i \odot c^{t-1} + b_i) \tag{2}$$

$$f^t = \sigma(\widehat{f}^t) = \sigma(W_f X^t + R_f h^{t-1} + p_f \odot c^{t-1} + b_f) \tag{3}$$

$$o^t = \sigma(\widehat{o}^t) = \sigma(W_o X^t + R_o h^{t-1} + p_o \odot c^t + b_o) \tag{4}$$

$$c^t = a^t \odot i^t + c^{t-1} \odot f^t \tag{5}$$

$$h^t = o^t \odot \tanh(c^t) \tag{6}$$

If N is the number of LSTM blocks and M is the number of inputs, then different weight used in the LSTM networks are:

- Input weights: $W_c, W_i, W_f, W_o \in^{N*M}$
- Recurrent weights: $R_c, R_i, R_f, R_o \in^{N*N}$
- Bias weights: $b_c, b_i, b_f, b_o \in^{N}$
- Peepholes weights: $p_i, p_f, p_o \in^{N}$.

where σ and ϕ denote the logistic sigmoid and hyperbolic tangent activation function. The term i, f, o, and c, respectively, denotes input gate, forget gate, output

gate, and cell activation vector. The element-wise multiplication is denoted by $\odot$ operation.

2.3 TDNN-LSTM Model

Better modeling power can be achieved by stacking more LSTM layers into the LSTM architecture. However, acoustic models with too many LSTM layers are hard to train and also have an issue in gradient. This problem can be handled by other variants of LSTM like highway LSTM [51] and residual LSTM [20, 52]. TDNN and LSTM models have their own pros and cons. Recently, better generalization capability was achieved by combining TDNN layers with LSTM layers [33]. The general combination of the convolution network with the recurrent neural network has been investigated in various works and proves their importance [2, 4, 33, 38]. The combination of TDNN with LSTM can be done in three ways:

1. Stacking LSTM over TDNN [2, 33]
2. Stacking TDNN over LSTM [2, 33]
3. Interleaving TDNN and LSTM [4, 33].

3 Multilingual Architecture for Hindi and Marathi Language

Multilingual ASR systems for low-resource languages have become more accepted in the past few years. In multilingual ASR systems, acoustic data of various languages have been shared which ensure greater coverage of contextual variation in all languages used in training [36]. This coverage is due to common phonemes shared by different languages. The acoustic data of different languages, known as source languages, are grouped together to form training data, and after training, the system has been tested for a particular language, which is known as a target language. In this work, Hindi and Marathi are used as source languages with approximately 7.5 h of training data with text transcription. In the recognition phase, decoding has been done by using the target language prior and posterior probabilities (Fig. 3).

Multilingual acoustic modeling can be done via two approaches, (i) universal phone set multilingual DNN, and (ii) shared-hidden layer multilingual DNN. In a universal phone set, a common phone set for all languages has been used. The common phone set can be obtained from IPA or via a data-driven approach [7]. In shared-hidden layer approach, hidden layers shared across all languages with language-specific output layer.

In this work, Marathi and Hindi language datasets have their own different phoneme set. We combined these phonemes set into one and train the acoustic model with this phoneme set and multilingual speech corpus. The architecture is shown in

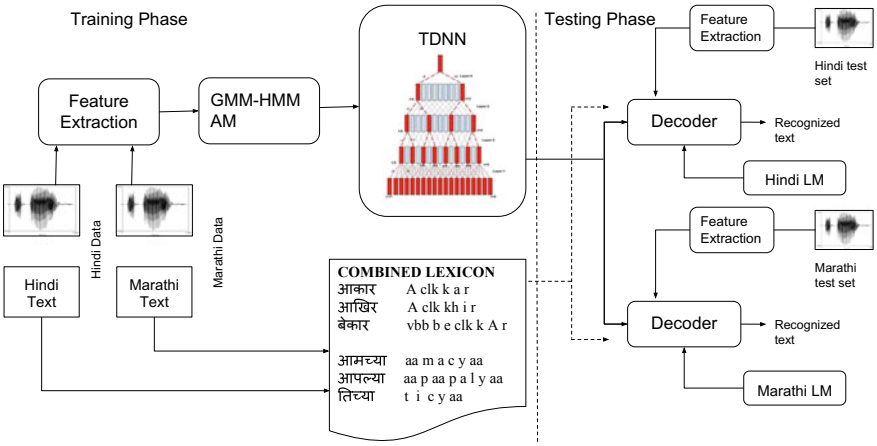


Fig. 3 Multilingual architecture of the ASR system for Hindi and Marathi language

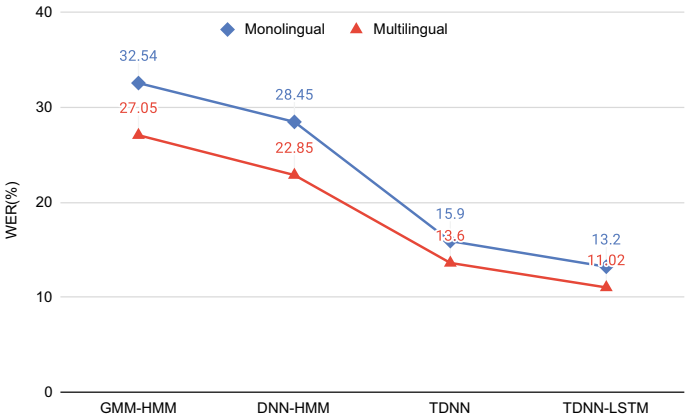


Fig. 4 Performance comparision of Hindi ASR system using different acoustic modeling techniques

Fig. 4. In such architecture, hidden layers act as a global feature extractor and cross-lingual knowledge transfer has been done in shared hidden layers. The input and hidden layers are common for all languages, but output layers are language specific to estimate the posterior probability of Senones of that language [7].

4 Corpus Details

For the Hindi language, we used a well-annotated, phonetically rich dataset [41] developed by TIFR, Mumbai in 2002. This dataset contains the speech of 100 speak-

Table 1 Details of Hindi and Marathi speech corpus

Languages	Training				Testing	Duration in Hours		
	No. of words	No. of utter-ances	No. of speakers		No. of utter-ances	Train data	Dev data	Total
			Male	Female				
Hindi	2345	800	55	25	200	2.1	0.16	2.5
Marathi	8819	5154	432	89	250	5.2	0.25	5.5

ers. The total duration of this speech dataset is 2.5 h, and it was recorded in a quiet room by two microphones on 16 kHz sampling frequency. The other details of the dataset can be found in Table 1. For the Marathi language, we used the speech corpus received from IIT, Guwahati.[1] This dataset is comparatively large with respect to the Hindi dataset.

Hindi dataset covers 2803 unique words, out of which 2345 words were used to train the acoustic model. In the case of the Marathi dataset, the size of vocabulary was 11701 unique words. The multilingual experiments were carried out using the common phone set, containing the phones of Hindi and Marathi language. The common phone set contains 108 phones for both languages, of which 29 were common. The monolingual Hindi ASR contains 68 phones, and 69 phones were used by Marathi monolingual ASR system.

5 Experimental Setup

All experiments were conducted on the dataset described in Sect. 4 using the Kaldi ASR toolkit [35]. The training and testing conditions remained the same throughout all experiments.

5.1 Language Modeling

The texts from the various sources were collected to train the monolingual language model. These sources were Emili corpus [8], magazines, newspapers (Hindi, Marathi), web text, and newsletters (Hindi, Marathi). For the Hindi language, 32k words were collected to train the language model. In the case of the Marathi language, 52k words were used. The SRI language modeling (SRILM) toolkit [44] was used to train the language model on the available text data of each language. The language models used in all experiments are Kneser-Nay smoothed tri-gram and 4-gram lan-

[1] Winter school on ASR, 16–28 May 2017, IIT Guwahati.

guage model. The interpolation of tri-gram and 4-gram (3G + 4G) language model is continued to further reduce the perplexity. The interpolated 3G + 4G language model has 320 perplexity on test transcription.

5.2 Acoustic Modeling

The multilingual acoustic modeling is carried out using the available training data of both languages. This work includes four different types of acoustic modeling described in sub-sections.

5.2.1 Gaussian Mixture Model-Hidden Markov Model (GMM-HMM)

After that tri-phone-based context-dependent acoustic modeling has been taken into consideration. In all experiments, we used 39-dimensional mel-frequency cepstral coefficient (MFCC) features including Δ and $\Delta - \Delta$ features. The baseline GMM-HMM system has 20k gaussian (approx 16 Gaussian per state) and 1250 cross-word tri-phone state for each language. For the multilingual GMM-HMM system, there are approximately 3000 cross-word tri-phone states with 50k Gaussians. The tri-phone-based GMM-HMM acoustic modeling was used to generate the alignment to train the neural network (NN) models.

5.2.2 DNN-HMM

DNN with 5 hidden layers and 1024 hidden units at each layer was used to train the acoustic model with 40-dimensional log-Mel filterbank (FBANK) features. DNN-HMM training was completed in four stages. First, the GMM-HMM model generates the alignment for DNN training. Then in the second phase, restricted Boltzmann machine (RBM) pre-training [8, 11] has been applied to initialize DNN. In the third stage, frame cross-entropy training has been done, and finally, optimization of cross-entropy objective function is done using state-level minimum Bayes risk (s-MBR) criterion [46]. The initial learning rate and mini-batch size chosen were 0.008 and 256, respectively. The time context size of 11 frames was used by concatenating +-5 frames.

5.2.3 Time-Delay Neural Network (TDNN)

TDNN is a feed-forward NN, able to effectively capture the long-term temporal dependencies [32]. TDNN system was trained with 40-dimensional hi-resolution MFCC features along with 100-dimensional i-vectors. In TDNN, speed perturbation has been applied to artificially increase the training data. This has been done by

the time-warping of raw audio [21]. First, the universal background model (UBM) system has been generated before TDNN training using all the available data. Then, i-vectors have been extracted. During training, all training data is available at the beginning, so the offline approach can be used for the computation of i-vectors. It is the other way round during testing. We use the standard recipe of the chain model given in the switchboard database in the Kaldi toolkit. The initial learning rate was chosen as 0.003000, which further decreases exponentially, and after 15 iterations, the final learning was stopped at 0.000150. The five-fold data augmentation has been used to train the TDNN acoustic model. The chain model is based on lattice-free maximum mutual information (LF-MMI) criterion that was used to train the model.

5.2.4 TDNN-LSTM

TDNN-LSTM (3 standards, 6 time-delay, and 3 LSTM layers) acoustic models were trained based on LF-MMI criterion [50]. For training, the standard recipe of the switchboard database given in the Kaldi toolkit was used. Along with 40-dimensional MFCC features, three-dimensional pitch features were also appended, as mentioned in the recipe. The training has been done using the parameters given in the recipe without additional parameter tuning. The five-fold data augmentation [21] and i-vector extraction was done as mentioned in the previous sub-section.

5.3 Pronunciation Lexicon

The pronunciation lexicon of the Hindi language contains 2803 words, and 13,650 words are in the Marathi language lexicon. The common lexicon contains 108 phones for both languages. Hindi language dataset used 68 phones, and Marathi contains 69 phones; out of these, 29 phones were shared by both languages. The final lexicon contains 14,558 words in total. All experiments in this work are based on closed vocabulary, which means test data contains no out-of-vocabulary words.

5.4 Tools and Performance Measures

The ASR system was implemented using the Kaldi speech recognition toolkit [35]. The switchboard recipe was used to extract the features, training, and decoding. For language modeling, the SRILM toolkit [44] was used. A detailed description of language modeling and acoustic modeling was given in previous sections. The ASR system performance was measured on WER.

6 Experimental Results and Discussion

In this section, we present and discuss the results of our work. First, we discuss the ASR system performance on the monolingual baseline system. Then in the next part, multilingual system results will be presented and discussed. The results are based on GMM-HMM, DNN-HMM, TDNN, and TDNN-LSTM-based acoustic modeling.

6.1 Training with One Language

This section shows the results of the experiments conducted with different acoustic modeling and language modeling. The switchboard recipes from the Kaldi toolkit were used to report the results. The results are shown in Table 2. It was observed that for both languages, the TDNN-LSTM acoustic modeling with interpolated (3G + 4G) language model showed lower WER compared to other system combinations. For Hindi, the WER was decreased from GMM-HMM (33%), DNN-HMM (29%), TDNN (18%) to TDNN-LSTM (16.2%) system. A similar trend was observed for the Marathi language. We tried to further reduce the WER by appending pitch features and interpolated language model. In this experiment, using the pitch features with interpolated language model shown improvements of 3% leading to WER of 13.2% for Hindi language and 37.95% for the Marathi language.

6.2 Training in Two Languages

This experiment highlights the fact that Indian languages are closely related to each other phonetically, and combining training data of other languages can improve the

Table 2 Experiment with one language using different acoustic and language modeling

S. No.	Acoustic Modeling	Features	Language Modeling	WER%	
				Hindi	Marathi
1	GMM-HMM	39-MFCC	3G	33.02	54.03
2	DNN-HMM	40-FBANK	3G	29.40	52.91
3	TDNN	40-MFCC	3G	18.40	44.45
4	TDNN	40-MFCC	3G + 4G	15.9	41.65
5	TDNN-LSTM	40-MFCC	3G	16.2	42.95
6	TDNN-LSTM	40-MFCC	3G + 4G	14.2	40.10
7	TDNN-LSTM	40-MFCC+ pitch	3G + 4G	13.2	37.95

Table 3 Comparision between monolingual system and multilingual system in WER(%) for the Hindi language (The multilingual system was trained with combined speech data of Hindi and Marathi languages, and the Hindi (3G + 4G) language model was used while testing.)

Acoustic Models	Features	WER(%)	
		Monolingual	Multilingual
GMM-HMM	MFCC-39	32.54	27.05
DNN-HMM	FBANK-40	28.45	22.85
TDNN	MFCC-40	15.90	13.60
TDNN-LSTM	MFCC-40+ pitch	13.20	11.02

Table 4 Comparision between monolingual system and multilingual system in WER(%) for Marathi language (The multilingual system was trained with combined speech data of Hindi and Marathi langauges, and the Marathi (3G + 4G) language model was used while testing)

Acoustic Models	Features	WER(%)	
		Monolingual	Multilingual
GMM-HMM	MFCC-39	53.02	45.65
DNN-HMM	FBANK-40	51.91	41.02
TDNN	MFCC-40	41.65	36.20
TDNN-LSTM	MFCC-40+ pitch	38.95	32.05

performance of the ASR system. It is known that Hindi and Marathi belong to the same Indo-Aryan family having similar phonetic sounds. Therefore, we train the ASR system using acoustic data of Hindi and Marathi language, and decoding is done by using the target language specific language model. The results of the experiments are shown in Table 3 & 4. It is evident that the WER of Hindi and Marathi decreases significantly by combining acoustic training data. The performance of the Hindi ASR system is improved. The WER reduced from 13.2% (for TDNN+LSTM as in Table 3) to 11.02%. A similar kind of results was reported by the Marathi ASR system.

As shown in Table 3, the WER for multilingual GMM-HMM system for Hindi with interpolated (3G + 4G) language model was recorded 27.05%, which was reported 32.54% in case of monolingual ASR system. Similar kinds of observations were found for the Marathi language as well. The overall improvement of different acoustic modeling over the monolingual system using a multilingual system can easily figure out in Table 4. This table clearly shows that combining similar acoustic data for training, further improves the performance of the ASR system.

7 Conclusion

We present a multilingual ASR system to recognize Hindi and Marathi language. Using two low-resource speech corpora, we evaluate the performance of the ASR system on different NN-based acoustic models. The combination of interpolated (3G + 4G) language modeling with TDNN-LSTM acoustic modeling performs better than other combinations. The best WER 11.02% and 32.05% have been recorded for Hindi and Marathi language, respectively, using TDNN-LSTM acoustic models. Training the acoustic models with the combined speech data of two languages resulted in significant reduction of WER in both languages. Thus, we conclude that multilingual ASR systems are the natural choice for low-resource Indian languages due to inter-language phonetic similarity. The experimental results also support the fact of high similarity between Hindi and Marathi language. Despite the fact that 15% relative improvement has been recorded by our best model over baseline, there is a need for further improvement. In the future, research area work can be focused on to study additional languages with more recent NN-based architectures.

References

1. Abraham B, Seeram T, Umesh S (2017) Transfer learning and distillation techniques to improve the acoustic modeling of low resource languages. In: INTERSPEECH, pp 2158–2162
2. Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E, Case C, Casper J, Catanzaro B, Cheng Q, Chen G et al (2016) Deep speech 2: End-to-end speech recognition in english and mandarin. In: International conference on machine learning, pp 173–182
3. Barker J, Watanabe S, Vincent E, Trmal J (2018) The fifth 'chime' speech separation and recognition challenge: dataset, task and baselines. arXiv preprint arXiv:1803.10609
4. Biswas A, de Wet F, van der Westhuizen E, Yilmaz E, Niesler T (2018) Multilingual neural network acoustic modelling for asr of under-resourced english-isizulu code-switched speech. In: INTERSPEECH, pp 2603–2607
5. Chellapriyadharshini M, Toffy A, Ramasubramanian V et al (2018) Semi-supervised and active-learning scenarios: efficient acoustic model refinement for a low resource Indian language. arXiv preprint arXiv:1810.06635
6. Chu SM, Povey D, Kuo HK, Mangu L, Zhang S, Shi Q, Qin Y (2010) The 2009 ibm gale mandarin broadcast transcription system. In: 2010 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 4374–4377
7. Cui J, Kingsbury B, Ramabhadran B, Sethy A, Audhkhasi K, Cui X, Kislal E, Mangu L, Nussbaum-Thom M, Picheny M et al (2015) Multilingual representations for low resource speech recognition and keyword search. In: 2015 IEEE workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE, pp 259–266
8. Dahl GE, Yu D, Deng L, Acero A (2011) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Language Process 20(1):30–42
9. Dash D, Kim MJ, Teplansky K, Wang J (2018) Automatic speech recognition with articulatory information and a unified dictionary for Hindi, Marathi, Bengali and Oriya. In: INTERSPEECH, pp 1046–1050
10. Dua M, Aggarwal RK, Biswas M (2019) Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modeling. Neural Comput Appl 31(10):6747–6755

11. Eastwood M, Jayne C (2013) Restricted Boltzmann machines for pre-training deep Gaussian networks. In: The 2013 International Joint Conference on Neural Networks (IJCNN). IEEE, pp 1–8
12. Fathima N, Patel T, Mahima C, Iyengar A (2018) Tdnn-based multilingual speech recognition system for low resource Indian languages. In: INTERSPEECH, pp 3197–3201
13. Gales MJ, Knill KM, Ragni A, Rath SP (2014) Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In: Fourth international workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014), pp 16–23. International Speech Communication Association (ISCA)
14. Hain T, Woodland P, Evermann G, Povey D (2000) The cu-htk march 2000 hub5e transcription system. In: Proceedings of speech transcription workshop. vol 1, Citeseer
15. Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A et al (2014) Deep speech: scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567
16. Hernandez F, Nguyen V, Ghannay S, Tomashenko N, Estève Y (2018) Ted-lium 3: twice as much data and corpus repartition for experiments on speaker adaptation. In: International conference on speech and computer. Springer, Berlin, pp 198–208
17. Huang JT, Li J, Yu D, Deng L, Gong Y (2013) Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 7304–7308
18. Kadyan V, Mantri A, Aggarwal R (2018) Refinement of hmm model parameters for Punjabi automatic speech recognition (pasr) system. IETE J Res 64(5):673–688
19. Karáfidt M, Baskar MK, Veselỳ K, Grézl F, Burget L, Černockỳ J (2018) Analysis of multilingual blstm acoustic model on low and high resource languages. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5789–5793
20. Kim J, El-Khamy M, Lee J (2017) Residual lstm: design of a deep recurrent architecture for distant speech recognition. arXiv preprint arXiv:1701.03360
21. Ko T, Peddinti V, Povey D, Khudanpur S (2015) Audio augmentation for speech recognition. In: Sixteenth annual conference of the international speech communication association
22. Kreyssig FL, Zhang C, Woodland PC (2018) Improved tdnns using deep kernels and frequency dependent grid-rnns. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 4864–4868
23. Kumar A, Aggarwal R (2020) Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modeling. J Intelligent Syst 30(1):165–179
24. Kumar A, Aggarwal R (2020) A time delay neural network acoustic modeling for Hindi speech recognition. In: Advances in data and information sciences. Springer, Berlin, pp 425–432
25. Kumar A, Aggarwal RK (2020) Hindi speech recognition using time delay neural network acoustic modeling with i-vector adaptation. Int J Speech Technol 1–12
26. Kumar A, Aggarwal RK (2020) A hybrid cnn-ligru acoustic modeling using raw waveform sincnet for hindi asr. Comput Sci 21(4)
27. Lang KJ, Waibel AH, Hinton GE (1990) A time-delay neural network architecture for isolated word recognition. Neural Netw 3(1):23–43
28. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5206–5210
29. Passricha V, Aggarwal RK (2019) Convolutional support vector machines for speech recognition. Int J Speech Technol 22(3):601–609
30. Peddinti V, Chen G, Manohar V, Ko T, Povey D, Khudanpur S (2015) Jhu aspire system: robust lvcsr with tdnns, ivector adaptation and rnn-lms. In: ASRU, pp 539–546
31. Peddinti V, Chen G, Povey D, Khudanpur S (2015) Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In: Sixteenth annual conference of the International Speech Communication Association

32. Peddinti V, Povey D, Khudanpur S (2015) A time delay neural network architecture for efficient modeling of long temporal contexts. In: Sixteenth annual conference of the International Speech Communication Association
33. Peddinti V, Wang Y, Povey D, Khudanpur S (2017) Low latency acoustic modeling using temporal convolution and lstms. IEEE Signal Process Lett 25(3):373–377
34. Povey D, Burget L, Agarwal M, Akyazi P, Kai F, Ghoshal A, Glembek O, Goel N, Karafiát M, Rastrow A et al (2011) The subspace gaussian mixture model-a structured model for speech recognition. Comput Speech Language 25(2):404–439
35. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P et al (2011) The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding, CONF. IEEE Signal Processing Society
36. Sahraeian R, Van Compernolle D (2018) Cross-entropy training of dnn ensemble acoustic models for low-resource asr. IEEE/ACM Trans Audio Speech Language Process 26(11):1991–2001
37. Sailor HB, Krishna MVS, Chhabra D, Patil AT, Kamble MR, Patil HA (2018) Da-iict/iiitv system for low resource speech recognition challenge 2018. In: INTERSPEECH, pp 3187–3191
38. Sainath TN, Vinyals O, Senior A, Sak H (2015) Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 4580–4584
39. Sak H, Senior A, Beaufays F (2014) Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. arXiv preprint arXiv:1402.1128
40. Sak H, Senior AW, Beaufays F (2014) Long short-term memory recurrent neural network architectures for large scale acoustic modeling
41. Samudravijaya K, Rao P, Agrawal S (2000) Hindi speech database. In: Sixth international conference on spoken language processing
42. Schalkwyk J, Beeferman D, Beaufays F, Byrne B, Chelba C, Cohen M, Kamvar M, Strope B (2010) Your word is my command: Google search by voice: a case study. In: Advances in speech recognition. Springer, Berlin, pp 61–90
43. Shetty VM, Sharon RA, Abraham B, Seeram T, Prakash A, Ravi N, Umesh S (2018) Articulatory and stacked bottleneck features for low resource speech recognition. In: INTERSPEECH, pp 3202–3206
44. Stolcke A (2002) Srilm-an extensible language modeling toolkit. In: Seventh international conference on spoken language processing
45. Tüske Z, Pinto J, Willett D, Schlüter R (2013) Investigation on cross-and multilingual mlp features under matched and mismatched acoustical conditions. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, pp 7349–7353
46. Veselỳ K, Ghoshal A, Burget L, Povey D (2013) Sequence-discriminative training of deep neural networks. Interspeech 2013:2345–2349
47. Veselỳ K, Karafiát M, Grézl F, Janda M, Egorova E (2012) The language-independent bottleneck features. In: 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp 336–341
48. Vu NT, Schultz T (2013) Multilingual multilayer perceptron for rapid language adaptation between and across language families. In: Interspeech, pp 515–519
49. Vydana HK, Gurugubelli K, Vegesna VVR, Vuppala AK (2018) An exploration towards joint acoustic modeling for indian languages: Iiit-h submission for low resource speech recognition challenge for Indian languages, interspeech 2018. In: INTERSPEECH, pp 3192–3196
50. Yu D, Li J (2017) Recent progresses in deep learning based acoustic models. IEEE/CAA J Automatica Sinica 4(3):396–409
51. Zhang Y, Chen G, Yu D, Yaco K, Khudanpur S, Glass J (2016) Highway long short-term memory rnns for distant speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 5755–5759
52. Zhao Y, Xu S, Xu B (2016) Multidimensional residual learning based on recurrent neural networks for acoustic modeling. In: Interspeech, pp 3419–3423

Vowel-Based Acoustic and Prosodic Study of Three Manipuri Dialects

Thangjam Clarinda Devi, Leihaorambam Sarbajit Singh, and Kabita Thaoroijam

Abstract The analysis of segmental and supra-segmental characteristics provides vital cues to dialectal variation of a language. There has been no comparative study on acoustic and prosodic features of the dialects of Manipuri. In this paper, we present a description of dialectal variation for three of the major dialects of Manipuri, namely Imphal, Kakching and Sekmai, based on the observation and analysis of speech data. The Imphal dialect spoken by the speakers, who live in the capital city of Manipur state known as Imphal, is taken to be the standard form for Manipuri language. Kakching and Sekmai dialect speakers live in Kakching and Awang Sekmai, respectively. Formant frequency, segment duration, energy values and pitch values are measured to study the vowel characteristics. The findings of the study show that dialectal variation is well represented by acoustic and prosodic variations in the three dialects. This preliminary analysis will be helpful for a more detailed comparative study of Manipuri dialects in the future.

Keywords Manipuri dialects · Vowels · Formant frequencies · Prosodic features

1 Introduction

Dialects of a language are identified by the unique speaking patterns followed by the community of speakers belonging to a particular geographical region. Dialects mainly exhibit grammatical, phonological and prosodic differences among them [1]. One of the primary causes of speech variabilities is dialects and is responsible for degradation in the performance of automatic speech recognition (ASR) [2]. The accuracy of an ASR can be improved by including the knowledge about dialects in

T. Clarinda Devi · L. Sarbajit Singh · K. Thaoroijam (✉)
Indian Institute of Information Technology Senapati, Manipur Senapati, India
e-mail: kabita@iiitmanipur.ac.in

T. Clarinda Devi
e-mail: thangc@iiitmanipur.ac.in

L. Sarbajit Singh
e-mail: sarbajit@iiitmanipur.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_34

pronunciation and acoustic model. The study of the effect of dialects to obtain ASR robustness is a major research topic.

Dialect specific information in speech signal is contained at both the segmental and supra-segmental levels. By observing the shapes of vocal tract during the production of a sound, information specific to spoken dialect can be obtained. This shape can be described and distinguished by the spectral envelope, and values can be used at segmental level to represent the information. Segment duration, energy values and pitch contour can be useful in giving prosodic information to study dialectal influences at supra-segmental level [3].

A few studies on the analysis of dialects of Manipuri language have been carried out in the recent past [4–7]. But, most of the works are descriptive in nature, and no previous attempts have been made in vowel-based acoustic and prosodic study of Manipuri dialects. Unavailability of such experimental studies in Manipuri language has been the motivation for the current work. Manipuri is a Tibeto-Burman sub-group of Sino-Tibetan language family. It is the official language of the state of Manipur. It was included in the Eighth Schedule to the Constitution of India on 20 August 1992 [8]. Manipuri has dialects of Imphal, Kakching, Sekmai, Andro, Khurkhul, Phayeng, Kwatha, Thanga, etc. Imphal dialect is considered the standard form of speech. The forms of Manipuri speech found in Tripura, Myanmar and Bangladesh still show clear dialectal variation [6]. The present study focuses only on three Manipuri dialects, namely Imphal, Kakching and Sekmai. The Imphal dialect, spoken around the capital city, is considered to be the standard variety. The speakers of Kakching dialect live in Kakching, a district which is situated to the south of Imphal valley at a distance of about 45 kms. The speakers of Sekmai dialect live in Awang Sekmai, situated near Imphal–Dimapur road about 19 kms from Imphal in the North. According to phoneme inventory, Manipuri has 38 phonemes, 36 segmental and 2 supra-segmental phonemes. Out of these 38 phonemes, 24 are consonants, 6 are monophthongs, 6 are diphthongs and the remaining 2 are tones. It is agreed that the language has 15 (/p/, /p^h/, /t/, /t^h/, /c/, /k/, /k^h/, /m/, /n/, /ŋ/, /s/, /h/, /l/, /w/ and /j/) indigenous consonants and the remaining 9 (/b/, /b^h/, /d/, /d^h/, /g/, /g^h/, /ɟ/, /$ɟ^h$/ and /ɹ/) are exotic consonants. The six monophthongs are /i/, /e/, /ə/, /a/, /u/ and /o/, and the six diphthongs are /əi/, /əu/, /ai/, /au/, /oi/ and /ui/ [9].

Rest of the paper is organized with the following sections: the description of the Manipuri vowel dialect dataset is discussed in Sect. 2. Feature extraction and preliminary analysis are given in Sect. 3. Section 4 contains conclusion and scope for future works.

2 The Dialect Dataset

Manipuri Speech Corpora collected from Linguistic Data Consortium for Indian Languages (LDC-IL) [10, 11] is used in this study. The data content chosen is S: SENTENCES. To ensure that all the types of syntactic structures are covered in

the speech data, a set of sentences have been constructed with the help of language experts and linguists. The words used in these sentences are common words which are found in day-to-day life. The sentences consist of words where the phonemes can occur in various positions and are meant to show the dialectal variants of the speakers. The dataset contains speech files from 150 speakers, and each speaker has uttered 25 randomly selected sentences out of 208 unique sentences. The age of the speakers varies from 16 to 51 years. All the speakers were literate enough and could read the textual sentences provided to them.

2.1 Annotation of Data

Manual annotation at phoneme level is done to identify the steady-state vowel regions of all the vowels present in the dataset using Praat speech analyser toolkit [12]. An example of an annotation is shown in Fig. 1. Annotation consisted of two tiers, namely phoneme and sentence tiers.

2.2 Vowel Frequency Distribution

For the preparation of the vowel dialect dataset, only a subset of the speech corpora is used. Brief details of dataset used in this study is shown in Table 1. A total of 1080 vowels have been analysed from the sentences uttered by forty-five speakers. Almost equal distribution of gender in each dialect has been taken into consideration.

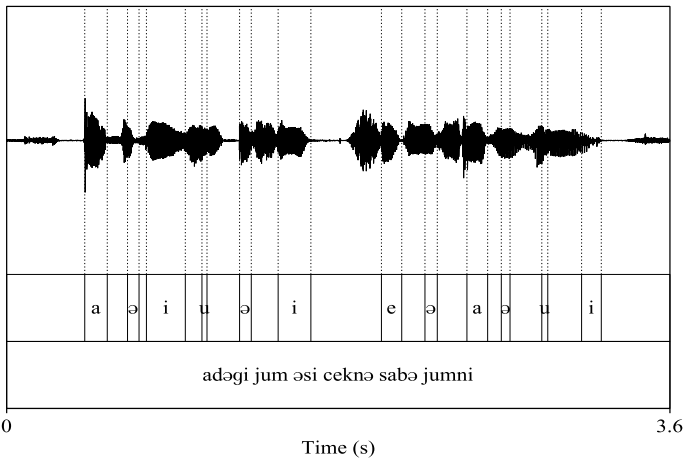


Fig. 1 Two-tier manual annotation of the speech data using Praat

Table 1 Vowel dataset details

Sl. no.	Dialect	Total no. of vowels	No. of speakers (M + F)
1	Imphal	320	7 + 8
2	Kakching	370	8 + 7
3	Sekmai	390	7 + 8

The table shows the description of the vowel dataset (M: Male, F: Female)

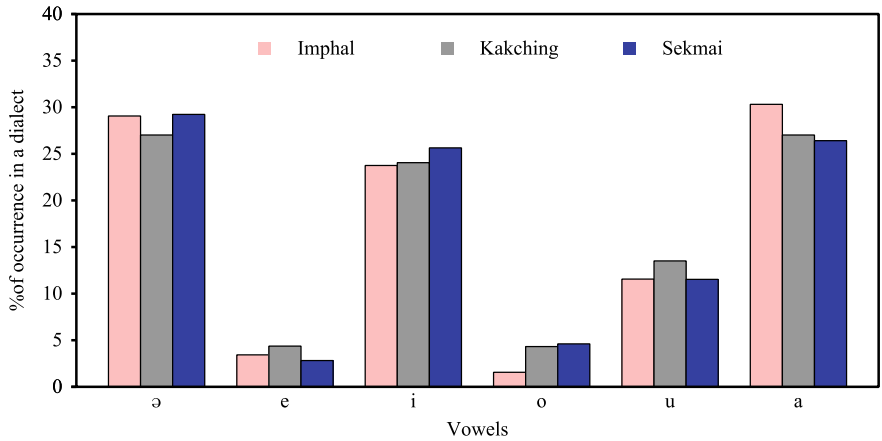


Fig. 2 Percentage of occurrence of each vowel out of a total of 320, 370 and 390 vowels in Imphal, Kakching and Sekmai, respectively

The frequency distribution of the vowels belonging to the three dialects is shown below in Fig. 2. There are six vowels, i.e. /ə/, /e/, /i/, /o/, /u/ and /a/ occurring in all the three dialects. Even some diphthongs also exist in the dataset, but only monophthongs are considered for the present study. The most frequent vowel is /a/ in Imphal dialect, both /ə/ and /a/ in Kakching dialect and /ə/ in Sekmai dialect.

3 Feature Extraction and Preliminary Analysis

In order to objectively differentiate the vowels in the three dialects, acoustic-phonetic and prosodic features at segmental and supra-segmental levels, respectively, are examined by using Praat toolkit. Formant frequencies, namely F1 and F2, are extracted as acoustic features from vowel steady regions which give us the spectral information at the segmental level. Segment duration, energy values and pitch contour values are analysed for the prosodic information at supra-segmental level [3]. The extraction process and the preliminary analysis are discussed in the below sections.

3.1 Formant Analysis

Figure 3 shows a traditional F2-F1 (in Hertz) plot, drawn using mean values of F2 and F1 of all six vowels from three dialects. Plot shows the vowels in the order /i/, /e/, /ə/, /a/, /o/, and /u/. While F1 gives information about vowel height, F2 provides information on vowel front or back position. Both F1 and F2 frequencies of the vowels are calculated at vowel midpoint.

From the figure, it is observed that the vowel space is larger for Imphal dialect in comparison with the other two dialects. The front vowels /i/ and /e/ of Imphal dialect have higher F2 value than the other two dialects. Central-low vowel /ə/ is found with almost similar F1 and F2 values across all dialects. Vowel /u/ of Imphal dialect has the lowest F2 and F1 values among the three dialects. Back-mid vowel /o/ from Sekmai dialect seems to have the higher F1 and F2 values compared to the other dialects.

The plot shows the existence of significant variation in vowel pronunciation among the three Manipuri dialects.

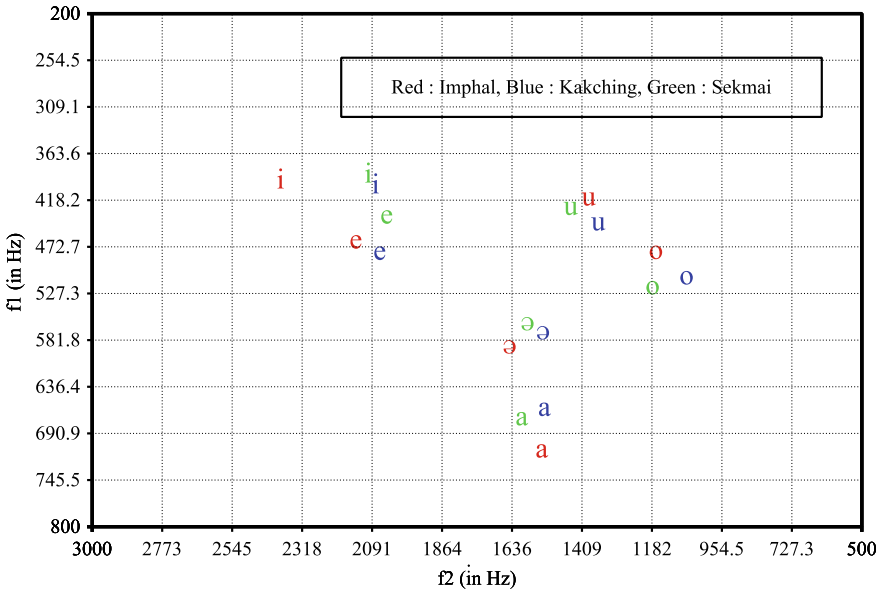


Fig. 3 Average vowel plots of three dialects with F2 and F1

3.2 Segment Duration

Dialects can also be discriminated on the basis of the length of the segment. The length is mainly concerned with vowel duration used in the segment [13]. This prosodic feature represents the perceptual property like speaking rate of speakers from the three dialects. Mean vowel duration of the six vowels when spoken by speakers of three different dialects is presented as a bar graph in Fig. 4. The time taken by Sekmai speakers in producing vowel sounds like /ə/, /u/ and /a/ is comparatively higher than the other speakers from different dialects. It can be further observed that Kakching speakers have higher speaking rate than other speakers from different dialects.

The bar graph clearly shows that the vowel qualities of these dialects are distinct from each other.

3.3 Energy Analysis

Energy feature associated with the speech signal is time varying in nature. This prosodic feature corresponds to the perceptual property, i.e. loudness. In this study, short-time energy is computed from vowel sounds for describing the loudness. It plays a significant role in human perception. Energy variations with time are measured concerning samples amplitudes within a frame [14].

Using the extracted energy values of all the six vowels in each dialect, a box plot is drawn providing a simple graphical summary of data in Fig. 5. The statistical information, namely mean, median, max, min, lower and upper quartiles, has been

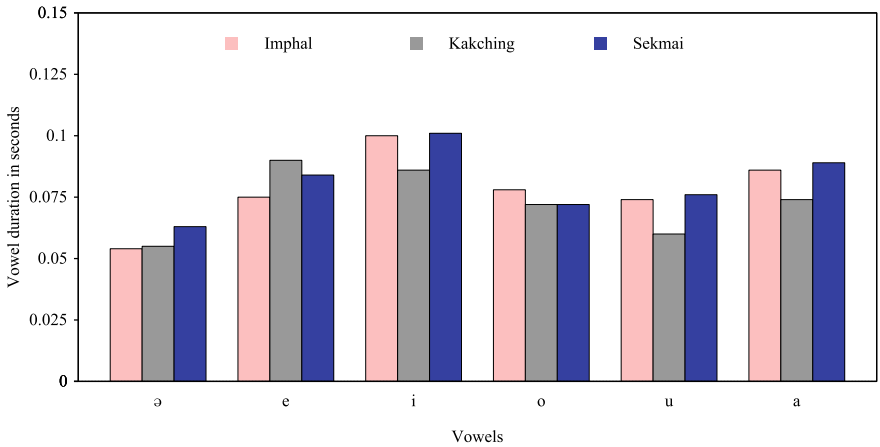


Fig. 4 Comparative vowel duration of Manipuri dialects

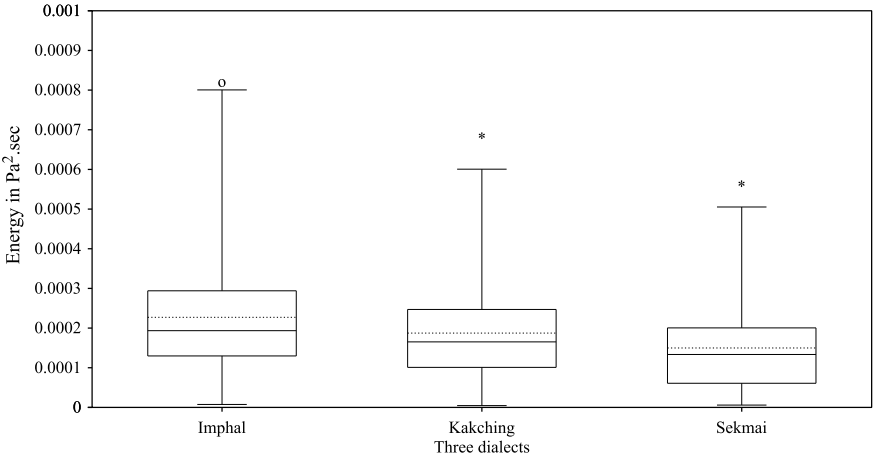


Fig. 5 Energy statistics for the speakers of three dialects

analysed. The horizontal lines of the rectangle corresponds to lower, median and upper quartiles, respectively. The dotted line corresponds to the mean, and "o" and "*" are the outliers.

It is observed that inter-quartile range is wider for Imphal dialect indicating that these speakers use high energy and have a large varying energy pattern. Range is narrower in Sekmai dialect as these speakers follow lower energy ranges and use a similar speaking style across the region. Outliers beyond upper inner fence are represented with "*" found in both Kakching and Sekmai dialects, and "o" represents the outliers beyond upper outer fence found only in Imphal dialect. All dialect outliers are skewed towards maximum value.

3.4 Pitch Analysis

In this section, the fundamental frequency (F0) values of the six vowels are examined. Figure 6 shows the bar plot of average F0 values (in Hz) of the six vowels in three Manipuri dialects, and Fig. 7. shows the bar plot of gender-dependent average F0 values (in Hz) of three Manipuri dialects. It is observed that the F0 values differs from dialect to dialect so we assume that F0 values are influenced by dialectal variation. Imphal dialect speakers have higher F0 values of vowels except /e/ compared to other dialects. Further cross-dialectal study is required for in-depth analysis.

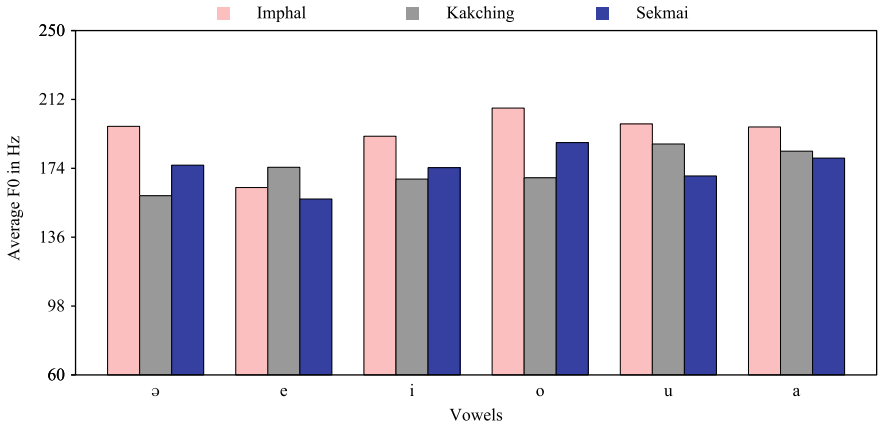


Fig. 6 Comparative vowel average F0 values of Manipuri dialects

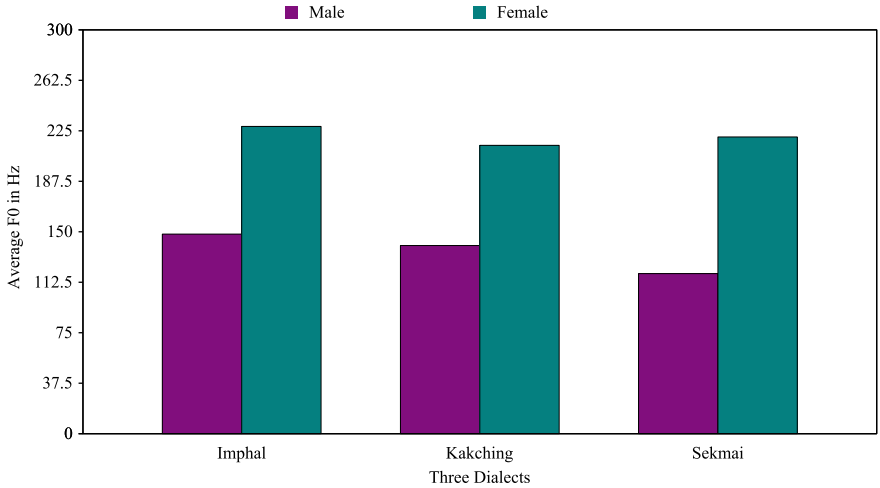


Fig. 7 Comparative gender-dependent average F0 values of Manipuri dialects

4 Conclusion and Future Work

This study has reported the vowel-based acoustic and prosodic analysis of three Manipuri dialects after the creation of vowel dialect dataset from the Manipuri speech corpora. Three major dialects, namely Imphal, Kakching and Sekmai have been considered for the analysis. Initial analysis shows a difference in the vowel spaces of the three dialects. Prosodic analysis showed that significant difference is present in the dialect varieties. This analysis gives preliminary results, and more cross-dialectal study of stress and intonation variations would help for in-depth analysis. Extensive work with larger dataset is currently under process which would probably show better

results and make the analysis more conclusive. The result of the analysis can further be used in the application of automatic dialect identification for Manipuri language.

Acknowledgements The authors would like to express sincere gratitude towards TEQIP-III, Indian Institute of Information Technology Senapati, Manipur, for providing the financial assistance for this study. Also, our sincere and heartfelt thanks to the whole team of Linguistic Data Consortium for Indian Languages (LDC-IL) and Central Institute of Indian Languages, Mysore, for creating the Manipuri speech corpora and allowing us to use it for this analysis.

References

1. Chambers JK, Trudgill P (1998) Dialectology. Cambridge University Press, Cambridge
2. Harris MJ, Gries ST, Miglio VG (2014) Prosody and its application to forensic linguistics. In: Linguistic evidence in security, law and intelligence, 19, vol 2, issue 2
3. Sinha S, Jain A, Agrawal SS (2014) Speech processing for Hindi dialect recognition. In: Advances in signal processing and intelligent recognition systems. Springer, Cham, pp 161–169
4. Devi KC (2018) Vowels in the phonology of Phayeng dialects. ZENITH Int J Multidisc Res 8(7):49–58
5. Haokip P (2011) The languages of Manipur: a case study of the Kuki-Chin languages. Linguistics of the Tibeto-Burman Area
6. Devi CC (2016) A comparative study of Imphal. Sekmai and Khurkhul dialects of Meiteirol, Shodhganga https://shodhganga.inflibnet.ac.in/handle/10603/105254
7. Devi LM (2016) A comparative study of Imphal, Andro. Koutruk and Kakching dialects of Meiteiron, Shodhganga https://shodhganga.inflibnet.ac.in/handle/10603/105256
8. Shodhganga homepage. https://shodhganga.inflibnet.ac.in/handle/10603/105256
9. Singh LS, Thaoroijam K, Das PK (2007) Written Manipuri (Meiteiron)-from phoneme to grapheme. In: Language In India 76
10. Ramamoorthy L, Choudhary N, Meetei AN, Chanu YP, Singh LA (2019) Manipuri raw speech corpus. Central Institute of Indian Languages, Mysore http://www.ldcil.org/resourcesSpeechCorp.aspx
11. Choudhary N, Rajesh N, Manasa G, Ramamoorthy L (2019) LDC-IL raw speech corpora: an overview. In: Linguistic resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore, pp 160–174
12. Boersma P, Weenink D (1992) Praat: a system for doing phonetics by computer
13. Sinha S (2015) Analysis and recognition of dialects of Hindi speech. Int J Sci Res Comput Sci Eng 3(5)
14. Chittaragi NB, Koolagudi SG (2019) Acoustic-phonetic feature based Kannada dialect identification from vowel sounds. Int J Speech Technol (Springer)

A Continuous Speech Recognition System for Bangla Language

Rajat Paul and K. Samudravijaya

Abstract This paper presents a Bangla continuous speech recognition system implemented using a publicly available speech corpus collected using the crowd-sourcing method. The open-source Kaldi speech recognition toolkit was used to train different acoustic models. Various language models were trained using the IRSTLM software. A Bangla pronunciation dictionary was created following Indian language speech labels. The dictionary contains a sequence of phone labels corresponding to the pronunciation of words in the transcriptions of the speech data. A combination of various types of acoustics and language models was explored along with the optimal setting of some parameters. The best speech recognition system uses a subspace Gaussian mixture model in conjunction with the hidden Markov model as the acoustic model, and quadgram as the language model. The word error rate of the system using an adequately trained language model is as low as 5% in a threefold cross-validation experiment.

Keywords Automatic speech recognition · Bangla language · SGMM-HMM · Kaldi · ILSL12

1 Introduction

Voice interfaces with machines are becoming more popular these days. Speech recognition is an essential module of such interfaces. While automatic speech recognition (ASR) systems have been developed for the languages being spoken in the developed world, such work is gaining pace in the developing countries. Here, we present the implementation of an ASR system for the Bangla language using an open-source software toolkit and a publicly available, crowd-sourced speech corpus of Bangla.

R. Paul
Department of Computer Science, NERIM Group of Institution, Guwahati 781022, India

K. Samudravijaya (✉)
Centre for Linguistic Science and Technology, Indian Institute of Technology Guwahati, Guwahati 781039, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_35

1.1 Bangla Language

Bangla, also known as Bengali, is an eastern Indo-Aryan language with around 230 million speakers, mainly in Bangladesh, and in several states of India, i.e., West Bengal, Tripura, and Assam [1]. Bangla is the national language of Bangladesh. Bangla is one of the official languages of the Government of India.

1.2 Related Research Work

One of the initial experiments on ASR of recognition Bangla speech used Bangla sentences spoken by several male and female speakers of varying age groups [2]. The authors implemented an ASR system using Sphinx, a hidden Markov model (HMM)-based ASR toolkit. Subsequently, a continuous speech ASR system for Bangla was implemented by the same group [3]. A continuous Bangla ASR system based on deep neural network (DNN) in combination with HMM was implemented in 2019 by a group of researchers from Bangladesh [4]. The authors used an open-source Kaldi software toolkit [5] to train the ASR system. The word error rate (WER) of the DNN-HMM system was less than 1%. Bangla ASR systems based on convolutional neural network (CNN) as well as recurrent neural network (RNN) were implemented by a group of researchers in the year 2019 [6]. The CNN was trained with 30 samples of each word. The CNN-based ASR system had an average accuracy of 86%. The RNN-based system had a word recognition accuracy of about 80%.

The previous ASR systems of Bangla predominantly used read speech data recorded in a quiet environment. The ASR system reported here is implemented using speech data that was crowd sourced. Such a speech corpus would have a large number of variations. Consequently, recognizing such freestyle speech with high accuracy would be challenging. Yet, a system that can recognize such speech with low errors is a need to build practical applications. This paper presents research and development work carried out while implementing such a system using a publicly available speech database.

The rest of the paper is organized as follows. Section 2 deals with the details of spoken language resources used in this work. The experimental setup, various experiments carried out, and the results of these experiments are presented in Sect. 3. A discussion of the experimental observations and conclusions is given in Sect. 4.

2 Spoken Language Resources

In order to implement an ASR system using supervised training, spoken language resources are essential. Audio data, along with the corresponding transcriptions as a sequence of words spoken in the audio file, are the basic resources. If the ASR

system models linguistic units such as phones, a lexicon (pronunciation dictionary) is also needed. In this section, a description of the resources used in the current work is given.

Speech data in a given language can either be recorded in-house or such a corpus publicly available can be used. Implementing an ASR system that recognizes speech from a large number of people speaking in various contexts demands a large amount of speech data. Modern ASR systems for medium resourced languages are trained with speech data of size starting from hundreds of hours to several thousands of hours. It is difficult to collect such a large amount of speech data and generate corresponding word transcriptions. On the other hand, a few speech databases that are available to the public, for payment or free, are available for many languages. Open language archives community lists such public resources for many languages. Their Web page for Bangla language lists several text, lexical, and speech resources [7]. IARPA Babel Bangla Language Pack is a large (215 h) speech dataset that can be purchased by the public [8]. The speech of both conversational and read styles was recorded in narrowband speech format. In the recent past, the cost of Internet (data) connection has come down. So, people prefer conversation over VoIP channels rather than the traditional telephone (narrowband) channel. Hence, it is better to implement an ASR system that recognizes speech over popular channels. Recently, Google released such a wideband speech database for free use by the public; this is listed as the 53rd database at the open speech and language resources web site [9]. We used this Bangla speech data to train and test the Bangla ASR system. The speech corpus contained not only audio data but also its transcription. The next two sections describe the processing of textual and audio data carried out by us.

2.1 Text Corpus

The Bangla spoken language data, provided by Google Inc., contain audio files and their transcriptions in Bangla script [9]. Text corpus consists of 218,618 unique utterances, with a vocabulary of 55,644 unique words. The sentences are short. A sentence typically consists of two to five words. The transcriptions of the audio files are provided in a file called “utt_spk_text.tsv”. A line of this “tab separated value” file contains information about one audio file: fileID, anonymized UserID and the transcription of audio in the file. These three items in a row are delimited by one tab character. The fileId consists of 10 digits. The anonymized speakerId in the second column was composed of five alphanumeric characters. The third column contained the transcription of that audio in Bangla script.

2.2 Speech Data

The speech data comprises 16 compressed directories of wideband audio files in.flac (Free Lossless Audio Codec) format. The digitized samples of an audio file, in flac format, are compressed such that the original samples can be recovered exactly from the compressed flac file, without losing the quality of the audio. Each audio file was uncompressed and stored in the following format: 16000 Hz, 16 bit, mono, RIFF wav header. The dataset contains about 218,618 files; the total duration of the audio data is about 210 h.

The Bangla utterances are from both male and female speakers. However, during the anonymization of speaker information, even the gender identity of speakers is lost. We listened to one file corresponding to every speakerId and manually noted down the gender of the person. The speech corpus was recorded with 304 male speakers and 204 female speakers. This information will be used while dividing speech data into training and test subsets.

We wrote a perl script file that prefixed the speakerId to the fileId with an underscore between them. For example, an audio file, with fileId 00b1a91838, spoken by a speaker whose speakerId is 01e91 was renamed as 01e91_00b1a91838. This file naming convention helped us to organize information needed by the Kaldi toolkit in a convenient manner.

2.3 Pronunciation Dictionary

A pronunciation dictionary (lexicon) is the backbone of an ASR engine that models speech units such as phones. Pronunciation dictionary is the phonetic representation of the words which occur in the audio files. Each phone-like linguistic unit is represented by a label. Phoneticians prefer to use the International Phonetic Alphabet (IPA) designed by the International Phonetic Association [10]. It is easier to process labels by a computer if the labels are in ASCII format. Arpabet is an example of such a label set [11]. We followed the Indian Language Speech Labels (ILSL) convention [12, 13] designed by the ASR and TTS consortium of projects sponsored by the Ministry of Information and Communication Technology, Government of India.

The IPA and ILSL labels for the speech sounds of the Bangla language are shown in Table 1. In the first row of the table, the first column shows the IPA alphabet for the first sound in the Bangla alphabet: অ. The second column shows the ILSL symbol and the third column shows the sound in Bangla script. The phone অ, in spoken Bangla, is slightly different from the corresponding vowel in Hindi, for example. The tongue is retracted while pronouncing the vowel অ of Bangla. Hence, a special symbol “ax” is assigned to the “first” vowel in the Bangla alphabet. This is in contrast to the symbol “a” assigned to the first vowel of the Devanagari script used by the Hindi language.

Table 1 IPA and ILSL symbols of the sounds of Bangla language

IPA	ILSL	Bangla	IPA	ILSL	Bangla	IPA	ILSL	Bangla	IPA	ILSL	Bangla
ɔ	**ax**	অ	g	**g**	গ	ɳ	**nx**	ণ	m	**m**	ম
aː	**aa**	আ	gʰ	**gh**	ঘ	t̪	**t**	ত,ৎ	j	**y**	য়
ɪ, i	**i**	ই,ঈ	ŋ	**ng**	ঙ,ং	t̪ʰ	**th**	থ	r	**r**	র
u, ʊ	**u**	উ,ঊ	tʃ	**c**	চ	d̪	**d**	দ	l	**l**	ল
-	**rq**	ঋ,ৃ	tʃʰ	**ch**	ছ	d̪ʰ	**dh**	ধ	ʃ	**sh**	শ,ষ
e	**e**	এ	dʒ	**j**	জ,য	n	**n**	ন	s	**s**	স
oj	**oi**	ঐ	dʒʰ	**jh**	ঝ	p	**p**	প	ɦ	**h**	হ
o	**o**	ও	ʈ	**tx**	ট	pʰ	**ph**	ফ	z	**z**	ज़
oʊ	**ou**	ঔ	ʈʰ	**txh**	ঠ	b	**b**	ব	ɽ	**dxq**	ड़
k	**k**	ক	ɖ	**dx**	ড	bʰ	**bh**	ভ	ɽʰ	**dxhq**	ढ़
kʰ	**kh**	খ	ɖʰ	**dxh**	ঢ						

The first column contains the IPA symbol and the second column contains the ILSL label (in ASCII format) corresponding to the grapheme/phone on the third column

The freely available speech database does not contain a pronunciation dictionary. We had to create one. Since there is a nearly one-to-one correspondence between the Bangla script and spoken Bangla, a semi-automatic approach was taken to create the pronunciation dictionary. Starting from the set of transcriptions provided by the database, a sorted list of unique words was generated. A Python script was written that generated a sequence of ILSL labels corresponding to a grapheme of a word. This sequence was manually inspected and edited so that the label sequence matches with the sequence of Bangla phones while pronouncing that word.

The resultant lexicon (pronunciation dictionary) file has two columns separated by a tab. A small section of the lexicon is shown in Table 2. The first column contains the word in Bangla script. The second column contains the corresponding sequence of ILSL labels.

3 Experimental Results and Discussion

In this section, we describe the experimental setup, various experiments conducted to improve the performance of the ASR system and present a discussion of the results.

Table 2 A few exemplar lines of the Bangla pronunciation dictionary

Bangla words	Label sequence
অজিত	ax j i t
আশী	aa shi
উঠে	u txh e
একপাশে	e k p aa sh e
কমানো	k ax m aa n o

The right column contains the sequence of ILSL labels corresponding to the pronunciation of the word in the first column

3.1 Performance Evaluation Criterion

The ASR system can make three types of errors at the word level: deletion, insertion, and substitution. Word error rate (WER) is used as the measure of the performance of the ASR system. Lower WER, the better the system is. WER is defined the ratio of the sum of all three types of errors divided by the total number of words in the reference word transcription, converted to percent.

3.2 Baseline ASR System

We used Kaldi, the open-source toolkit for building ASR systems. We used the popular 39-dimensional feature vector comprising 13 mel-frequency cepstral coefficients (MFCCs) and their first and second time derivatives to represent short segments of 25 ms duration, with a frame shift of 10 ms.

Kaldi facilitates modeling the acoustic properties of both context-independent and context-dependent phones using HMMs. The IRSTLM toolkit is used to build n-gram language models. When each phone is modeled by one HMM, the resulting ASR system is denoted as a “monophone” (or simply “mono”) system. When an HMM models a phone whose left and right phonetic contexts are predecided, such a system is called a triphone-based ASR system. Kaldi, by default, permits the user to implement three types of triphone ASR systems based on the types of feature adaptation and speaker adaptation of the HMMs. These are known as tri1, tri2, and tir3. The tri3 is more sophisticated than tri2 and so on. A typical HMM has three “states” that represent subphonetic events. Each state of HMM is associated with a probability density function that describes, in a statistical sense, the acoustic properties of the subphonetic event. Generally, a Gaussian mixture model (GMM) is employed. When the probability density function associated with the states of an HMM is modeled by a GMM, such a system is called GMM-HMM. When the GMM is replaced by a subspace GMM (SGMM), the resulting system is known as SGMM-HMM. When a deep neural network (DNN) is used to estimate the posterior

Table 3 Statistics of the three subsets of data

Number of	text1	text2	text3	Total
Utterances	72,221	73,151	73,246	218,618
Speakers	170	169	169	508
Male speakers	92	107	105	304
Female speakers	78	62	64	204

probabilities of various (shared) states generating a given feature vector, the system is known as DNN-HMM.

When a SGMM or DNN is the acoustic model of an ASR system, training and testing the ASR system take a lot of time, especially because the database size is large. So, baseline ASR systems were implemented using the simpler four types (excluding SGMM and DNN) of acoustic models. A bigram language model was employed. The set of 218,618 files were divided into three nearly equal parts such that each part contains a nearly equal number of male and female speakers. Here, we used the gender identity of each speaker determined by manual listening of audio files.

The statistics of the three subsets of data are shown in Table 3. For example, the first subset, whose transcriptions are stored in a file called text1, contains 72,221 speech files, spoken by a set of 92 male and 78 female speakers. The first subset (text1) was designated as the test set. The acoustic models were trained with the speech data belonging to subsets 2 and 3. The bigram language model was trained with the transcriptions of the subsets 2 and 3 (text2 and text3).The WER (in %) of the ASR systems that used mono, tri1, tri2, and tri3 acoustic models were 49.2, 23.2, 24.8, and 22.8, respectively.

3.3 Effect of Language Model

The bigram model, used in the baseline system, uses the probabilities sequence of two words to disambiguate between acoustically similar (and hence confusable) words. The IRSTLM allows capturing linguistic constraints longer than two words. So, we retrained two additional language models that capture the probabilities of sequences of three and four words. We refer to these language models as trigram and quadgram models, respectively. The same set of test data was decoded using trigram and quadgram models. The resultant WER values for four types of acoustic models are shown in Table 4.

The decrease in WER with the employment of more sophisticated acoustic models and longer range linguistic constraints is shown in the form of a histogram in Fig. 1.

Table 4 WER (%) of ASR systems employing different combinations of four acoustic models and three language models

WER (%)				
Language model	Acoustic model			
	Mono	tri1	tri2	tri3
Bigram	49.2	23.2	24.8	22.8
Trigram	39.8	20.4	22.0	20.6
Quadgram	39.3	20.1	21.7	20.3

Not only does WER decrease with increasing details of acoustic modeling, but also decreases when the language model captures longer range syntactic constraints

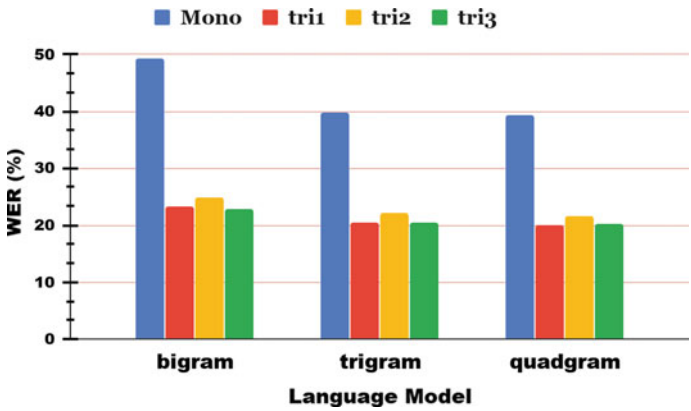


Fig. 1 Graphical representation of decreases in WER (%) of ASR systems using different language models (bigram, trigram, and quadgram) and using different acoustic models (mono, tri1, tri 2, and tri3). The ASR system that used the tri3 acoustic model and quadgram language model has the lowest WER of 20.1%

3.4 Optimal Number of Senones

Initial values of the number of Senones (shared HMM states), the language model weight and the word insertion penalty were 3000, 15 and 1.0 respectively. These values were used in the above experiments. We experimented with different values of these parameters. The WER values (%) for ASR systems using five acoustic models for three values of the number of Senones are shown in Table 5.

When the number of Senones was reduced from 3000 to 2000, there was a small but persistent decrease in the word error rate of tri2 and tri3 models. The overall error rate is lower when the number of Senones is 2000. For all future experiments, we set the number of Senones to 2000, word insertion penalty to zero and language model weight (lm_wt) to around 15 because they were the optimal values for this database.

Table 5 WER (%) of ASR systems employing five acoustic models as a function of three values of the number of Senones (shared HMM states)

WER (%)					
No. of Senones	Acoustic model				
	Mono	tri1	tri2	tri3	sgmm
3000	39.3	20.1	21.7	20.3	–
2000	39.4	20.6	21.3	20.0	17.1
1500	39.4	21.3	21.4	20.1	17.0

The WER was lower when the number of Senones was around 2000

3.5 Threefold Cross-Validation

The performance of the ASR system could be influenced by the quality of acoustic data in train and test subsets. To minimize such an effect, a k-fold cross-validation method was used for the evaluation of the performance of the ASR engine. We set $k = 3$. In the baseline ASR system, we used speech files and transcriptions of the first subset (text1) to test the performance of the system trained with the remaining data. We refer to this experiment as the fold1 experiment. One can designate data corresponding to text2 as the test data and evaluate the performance of the system trained with both text1 and text3 data subsets. This is a fold2 experiment. Similarly, a fold3 experiment can be conducted. The dataset of about 210 hours of audio was divided into three subsets such that each subset contains an approximately equal number of utterances as shown in Table 3. The arrangements of 3 subsets used for training and testing during threefold cross-validation are shown in Fig. 2. The subset (box) outlined in blue color is the test set. The other two subsets, shown as black

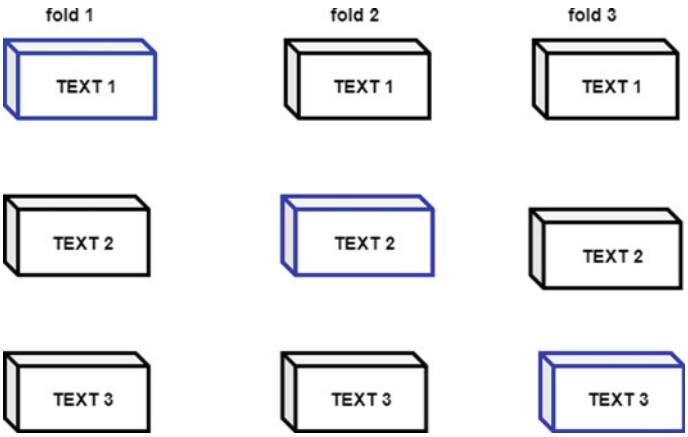


Fig. 2 Pictorial representation of the distribution of subsets for training and testing the ASR systems in threefold cross-validation experiments

Table 6 WER (%) of ASR systems employing five types of acoustic models in a threefold cross-validation experiment

	Acoustic model				
	Mono	tri1	tri2	tri3	sgmm
Fold1	37.5	20.6	21.4	20.3	16.7
Fold2	38.4	21.8	22.9	20.0	16.8
Fold3	37.8	20.8	21.7	19.4	16.3
Average	**37.9**	**21.1**	**22.0**	**19.9**	**16.6**

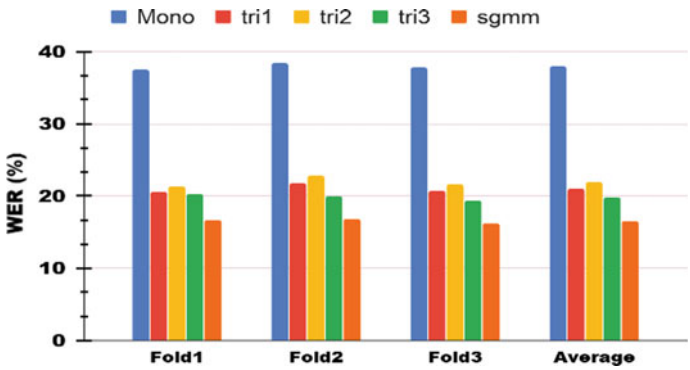


Fig. 3 Graphical representation of WER (%) of systems using different acoustic models (mono, tri1, tri2, tri3, and sgmm) in a threefold cross-validation experiment

boxes, are used to train the acoustic model. The language model is trained with the transcriptions of the training data.

Table 6 shows the word error rates of ASR systems using five types of acoustic models. The SGMM-HMM models show the lowest WER. The optimal language model weight (lm_wt) for ASR systems using mono, tri1, tri2, and sgmm acoustic models were 11, 15, 15, 17, and 14, respectively. The variation of WER as a function of acoustic models is shown in Fig. 3.

3.6 Language Model Training

An analysis of errors in the output of the ASR system revealed that many words in the test corpus did not occur in the transcriptions of the train data. Consequently, these words were not part of the language model that was trained solely from the transcriptions of the training data. Training of the language model can include any text, even outside the training data. Quite often, researchers crawl the web to accumulate text in the desired language and use a large amount of text data to train a language model that will be used during decoding. However, this also necessitates

generating a pronunciation dictionary for all the words in the large text corpora. We employed a shortcut that ensured that all the words in the test data are part of the language model. We trained the language model with not just the transcriptions of the training data but also with the transcriptions of the test dataset. We evaluated the performance of ASR systems decoded using this augmented language model.

Table 7 shows the WER (in %) of systems using the language model trained with transcriptions of both training and testing datasets. There is a significant reduction in WER values. The average WER reduces from 16.6% to 4.9%, i.e., by a factor of 3 or more, in case of SGMM-HMM-based ASR system. Such a significant reduction in WER values is graphically shown in Fig. 4.

Table 7 WER (%) of ASR systems employing five types of acoustic models in conjunction with a language model that is trained with transcriptions of the entire database

	Acoustic model				
	Mono	tri1	tri2	tri3	sgmm
Fold1	17.5	6.8	8.2	8.4	5.6
Fold2	16.6	5.6	7.5	6.7	4.6
Fold3	19.0	7.0	8.6	7.7	5.1
Average	**17.2**	**6.3**	**7.5**	**7.9**	**4.9**

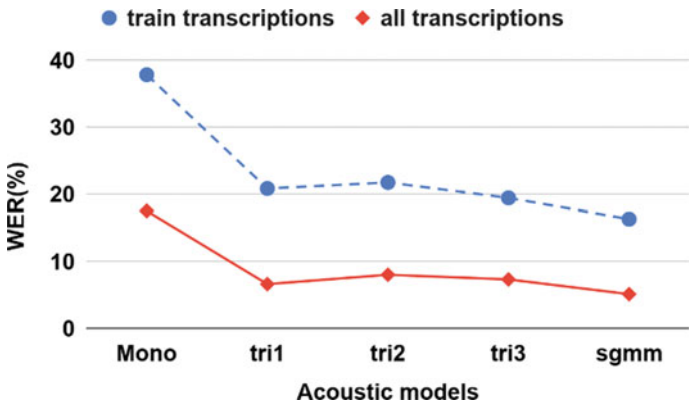


Fig. 4 Graphical representation of WER(%) of systems using two types of language models: either trained with transcriptions of only training data (blue filled circles connected by dashed lines) or trained with transcriptions of both training and test data (red diamonds connected by solid lines). The WER of SGMM-HMM system reduces by a factor or 3 by the usage of an adequately trained language model

4 Conclusions

The error rates of ASR systems reduce with the usage of sophisticated acoustic models such as SGMM-HMM. A DNN-HMM model is expected to perform in the task of recognizing casual speech with large vocabulary even better provided there is a large amount of data, of the order of 1000 h. However, the adequate training and testing of DNNs require huge computing power. We abandoned the training of DNNs for this reason at this moment.

As exemplified by Fig. 4, the language model plays a significant role in correct decoding of input speech even when the test speech contains four to five words. While the method we adopted to ensure that the language model contains all the words in the test dataset is imperfect, the WER observed sets an upper limit on the performance of the ASR system. Perhaps a better method is to include all the words of the test data, one word per line, in a temporary text file, and use it in addition to the training data transcriptions to train the language model.

Our hesitation to use a large amount of text, outside the training dataset, was the additional work load of preparing a large pronunciation dictionary. The recent end-to-end ASR systems obviate the need for such a lexicon. The output of such systems is at the character level, and they use character contexts rather than phonetic contexts. Probably, that is a direction to pursue in future.

References

1. Gordon R (2019) Ethnologue: languages of the world, 22nd edn. SIL International, Texas
2. Das B, Mandal S, Mitra P (2011) Bengali speech corpus for continuous automatic speech recognition system. In: 2011 international conference on speech database and assessments (Oriental COCOSDA), Hsinchu, pp 51–55. https://doi.org/10.1109/ICSDA.2011.6085979
3. Mandal S, Das B, Mitra P, Basu A (2011) A robust Bengali continuous speech recognizer using triphone and trigram language model. In: Aluru S et al (eds) Contemporary computing. IC3 2011. Communications in computer and information science, vol 168. Springer, Berlin. https://doi.org/10.1007/978-3-642-22606-9_61
4. Al Amin MA, Islam MT, Kibria S, Rahman MS (2019) Continuous Bengali speech recognition based on deep neural network. In: International conference on electrical, computer and communication engineering (ECCE)
5. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P et al (2011) The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding, no EPFL-CONF-192584. IEEE Signal Processing Society
6. Islam J, Mubassira M, Rakibul Islam M, Das AK (2019) A speech recognition system for Bengali language using recurrent neural network. In: IEEE 4th international conference on computer science and communication systems. https://www.researchgate.net/publication/335577021
7. OLAC resources in and about the Bengali language. https://olac.ldc.upenn.edu/language/ben
8. Bills A et al (2016) IARPA Babel Bengali language pack IARPA-babel103b-v0.4b LDC2016S08. Web Download. Linguistic Data Consortium, Philadelphia. https://catalog.ldc.upenn.edu/LDC2016S0

9. Kjartansson O, Sarin S, Pipatsrisawat K, Jansche M, Ha L (2018) Crowd-sourced speech corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In: Proceedings of the 6th international workshop on spoken language technologies for under-resourced languages, pp 52–55. https://doi.org/10.21437/SLTU.2018-11. https://openslr.org/53/
10. International Phonetic Association: the IPA chart. https://www.internationalphoneticassociation.org/content/ipa-chart
11. Klautau A (2001) ARPABET and the TIMIT alphabet. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf
12. Samudravijaya K, Murthy HA (2012) Indian language speech sound label set (ILSL12), 2012 developed by Indian Language TTS Consortium & ASR Consortium. https://www.iitm.ac.in/donlab/tts/downloads/cls/cls_v2.1.6.pdf
13. Samudravijaya K (2020) Indian language speech label (ILSL): a de facto national standard. Presented at International conference FRSM 2020, Silchar. https://www.iitg.ac.in/samudravijaya/tutorials/ILSL_V3.2.pdf

Indian Language Speech Label (ILSL): A De Facto National Standard

K. Samudravijaya

Abstract Speech can be viewed as a sequence of linguistic units such as phonemes. Traditional speech-to-text and text-to-speech systems are based on statistical models of acoustic–phonetic units such as phones. It is desirable to assign printable ASCII labels for acoustic–phonetic units to enable supervised training of models using a computer. Since the letters of the Roman alphabet are unable to label all speech sounds of Indian languages without ambiguity, a common set of labels for speech sounds of Indian languages was devised. This set, Indian Language Speech Labels, has been extensively used by major speech groups in India and has become a de facto standard. The design principles of this label set and the tables containing labels for the phones of 16 Indian languages are presented here.

Keywords Indian language · Speech label · Automatic speech recognition · ILSL

K. Samudravijaya (✉)
Centre for Linguistic Science and Technology, Indian Institute of Technology Guwahati, Guwahati 781039, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1_36

1 Introduction

Speech is a convenient medium of communication among human beings. It is one of the key traits of human species that enabled them to form settlements, which led to civilization. The primary goal of speech is delivery of a message through the medium of acoustic waves generated by human speech organs. Speech is spoken language. Thus, the speech signal has a well-defined mapping to a sequence of acoustic–phonetic units called phones. Writing systems were invented to visually represent verbal communication. A writing system is based on script comprising characters. There is a fuzzy mapping between phones and characters in Alphabetical, Abjad, and Abugida writing systems.

With advances in computer and communication technology, the human desire to interact with computers in the mode of her convenience has increased. Human–computer interaction in voice mode has become practical in many developed countries. Similar transformation in human–machine communication is happening in developing countries such as India. Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) technologies are the core modules of voice interfaces. The development of voice interfaces in indigenous languages demands implementation of ASR and TTS systems in local languages. Traditional ASR and TTS systems are based on statistical models of acoustic–phonetic units such as phones. It is desirable to have labels, preferably in terms of alphanumeric, in ASCII format, for acoustic–phonetic units to enable supervised training of models using a computer. This paper presents a description of such a labeling scheme for the acoustic–phonetic units of spoken Indian languages for the purpose of computer processing.

1.1 International Scene

The International Phonetic Alphabet (IPA) consists of more than 100 letters and diacritics to denote speech sounds used in languages of the world [1]. Several ASCII-based encoding systems of the IPA have been proposed. One of the earlier schemes is ARPAbet devised in the 1970s. It is a set of phonetic transcription codes developed by Advanced Research Projects Agency (ARPA) as a part of their Speech Understanding Research project. It represents phonemes and allophones of General American English with distinct sequences of ASCII characters [2]. ARPAbet is used in the CMU pronouncing dictionary. The well-known TIMIT speech database uses an extended version of the ARPAbet [3]. The Speech Assessment Methods Phonetic Alphabet (SAMPA) is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the IPA [4]. SAMPA was designed in the late 1980s, and used by many European languages. An extended version called X-SAMPA covered all characters of IPA. Consequently, the IPA was remapped into 7-bit ASCII. Based on X-SAMPA, a phonetic labeling convention for Standard Chinese, called SAMPA-C, was created to label spontaneous speech [5].

1.2 National Scene

With the advent of graphical terminals in the 1980s, standard codes for internal representation of text in Indian languages were evolved. This can be viewed as the task of assigning codes to the aksharas of the language [6]. In the context of multilingual computing, Graphic and Intelligence-based Script Technology (GIST) cards were developed by CDAC [7]. Later, Indian Script Code for Information Interchange (ISCII) evolved, and was published by the Bureau of Indian Standards in 1991 [8, 9]. ISCII is a coding scheme for representing various writing systems of India. It encoded Indic scripts and prescribed a Roman transliteration. These were useful for data entry and display of text in Indian scripts. Transliteration of the Indian scripts to the Roman script was achieved using diacritic marks. The National Library at Calcutta standardized the diacritic marks to be used for Romanization of Indian scripts, in 1988 [9].

For the development of spoken language technologies for Indian languages, a scheme for labeling speech sounds of Indian languages is needed. This is due to the fact that traditional speech-to-text and text-to-speech systems treat a speech signal as a sequence of linguistic units such as phones, and phone models are trained.

In the early 1990s, a multispeaker, general purpose speech database for Hindi Language was created. The speech files were annotated at the phonetic level and manually segmented. The labels were in ASCII alphanumeric format. Uppercase letters were used to denote long vowels and retroflex sounds [10]. Different speech research groups in the country followed homegrown labels for linguistic units while implementing speech systems. Some of these efforts were presented at national level conferences too [11]. When a national project in consortium mode for ASR [12] was initiated by the Technology Development in Indian Languages division of the Ministry of Electronics and Information Technology, the need for following a standard set of labels for Indian language speech units was felt. This gave rise to Indian Language Speech Label (ILSL) that was adopted by more than a dozen speech groups in India working not only in the field of ASR but also in the field of TTS [13, 14]. The salient features of ILSL will be presented in the next section.

2 Indian Language Speech Labels

Indian Language Speech Labels (ILSL) specifies a standard set of labels for speech sounds commonly used in Indian languages. The label set was initially designed to cater to 13 languages, processed by ASR or TTS consortia sponsored by the Government of India [12]. It was later extended to a few other languages spoken in and around India. These labels are meant to be used for computer processing of phones of spoken Indian languages. The scheme should not be confused with text to phoneme rules. The design of the labels for speech sounds followed a few simple principles.

1. Similar sounds in different languages are given a single label. This exploits the fact that most Indian scripts are derived from the Brahmi script.
2. The label set is designed such that the native script is largely recoverable from the transliteration while noting that the mapping is not one-to-one.

A label may consist of a sequence of alphanumeric characters of the Roman alphabet. A label does not contain any special character like quote, hyphen, etc., so that processing of labels by editing scripts such as perl is smooth. All labels are in lower case even though the labels are case insensitive. Case insensitivity ensures that the output of text processing is operating system agnostic. The labels for speech sounds will follow the popular convention as far as possible while noting that complete uniformity is not achievable.

There are variations in the transliteration schemes followed in different parts of the country. An outstanding example is the case of dental and retroflex sounds. The Roman script does not distinguish between the places of articulation of these two sets of plosives. The label "t" is used for unvoiced plosives of both places of articulation in most parts of the country. This is in contrast to the usage of "t" for unvoiced retroflex plosive and "th" for its dental counterpart in the southern part of India.

As far as possible, we will use Hindi, written in Devanagari script, to illustrate the conventions of ILSL. Since the number of speech sounds is larger than that of the letters of the Roman alphabet, a system of suffixes as well as letter combinations is used while labeling all speech sounds of Indian languages. The conventions followed while assigning labels to phones are described below.

2.1 Vowels

1. The label for a vowel matra is the same as that of the vowel.
2. Long vowels are indicated by the repetition of symbols of their short counterparts. For example, the label for the short vowel अ is "a" whereas the label for आ is "aa".
3. The label of a diphthong is generated as a concatenation of the labels of the corresponding vowels. The label for ऐ in Hindi is "ei". The exceptions to this rule are "ae", "ea", and "eu"; these are monophthongs.
4. *Nasalization* of a vowel: Use suffix n to denote nasalization of a vowel; e.g., "k a h aa" (कहा) versus "k a h aan" (कहाँ).
5. The *halant* in Indian scripts denotes the absence of the implicit "a" in Indian consonant characters. It is not a sound and hence there is no label for halant glyph. The morphological analyzer of the language shall delete the implicit "a" when a halant is present in the script.

2.2 Consonants

1. *Aspiration*: Use suffix h to denote aspiration: "k a" (क) versus "kh a" (ख).
2. *Retroflex* consonants: Use suffix x to denote retroflex place of articulation: "t a" (त) versus "tx a" (ट).
3. *Nukta*: Nukta (a filled dot below a consonant glyph) is a diacritic mark. It denotes a flap/tap or a fricative variant of the consonant. Use suffix "q" to denote the presence of a nukta: "dx a" (ड) versus "dxq a" (ड़).
4. *Anuswar* (a dot above a [vowel] glyph) indicates a nasal consonant sound. A common name is bindu. Bindu denotes a nasal after the vowel; the place of articulation of the nasal will be the same as that of the following consonant. If there is no consonant following the bindu, the vowel is nasalized.
5. *Geminated* sounds: The label for a geminated consonant is the label of the corresponding single consonant with the first letter of the label repeated. Example: "p a k aa" (पका) versus "p a kk aa" (पक्का) in Hindi. Additional examples: "a ddx aa" (अड्डा) in Hindi; "a ll a m" (**ginger**) in Tamil
6. Other *special cases*: Use suffix x to denote certain special cases: "ax" (अ) of Bangla; reduced vowel ("axx") in various languages; apical affricates of Marathi; special "r" of Dravidian languages etc.
7. Priority of suffixes: Some symbols may have multiple suffixes. In such cases the following is the priority (in decreasing order): x h q n
8. Punctuation marks: A "transliteration" module will retain the punctuation marks (exception: "।" and "॥" will be replaced by full stop); these are useful for prosody generation.

Tables 1, 2, and 3 list the symbols employed for speech sounds of 16 languages. Table 1 shows the symbols for vowels. Table 2 shows symbols for plosives and nasals. Table 3 lists the ILSL symbols for other consonants. Whenever a phone does not occur in a language, the corresponding cell contains a dash character.

An IPA symbol in the third column of the tables refers to a sound in an exemplar (Hindi/Tamil/other) language. The correspondence between the IPA symbols in a row with the glyphs of all languages is approximate since an ILSL label represents phonetically similar (but not necessarily exactly equal) sounds of Indian languages.

Figure 1 shows the ILSL symbols associated with the Devanagari alphabet; complex ligatures of the script are not shown for the sake of simplicity here. The "a" implicit in the glyph of the consonants is not shown here for the sake of simplicity.

2.3 Language-Specific Notes

This subsection has notes on sounds (and labels) specific to a subset of the languages. Whenever possible, minimal pairs for language-specific phonemes are provided. In other cases, exemplar words containing the specific phone are written.

Table 1 ILSL and IPA symbols for vowels of several languages of the Indian subcontinent

Sr.No.	ILSL	IPA	Hindi	Marathi	Rajasthani	Gujarati	Bodo	Tamil	Malayalam	Telugu	Kannada	Bengla	Assamese	Manipuri	Dzongkha	Galo
1	a	a	अ	अ	अ	અ	अ	அ	അ	అ	ಅ		–	অ	অ	a
2	ax	ɔ	–	ऑ	–	ઑ	–	–	–	–	–	অ	অ	–		
3	aa	aː	आ	आ	आ	આ	आ	ஆ	ആ	ఆ	ಆ	আ	আ	আ		
4	axx	ə	–	–	–	–	–	–	–	–	–	–	–	অ		v
5	i	ɪ, i	इ	इ	इ	ઇ	इ	இ	ഇ	ఇ	ಇ	ই	ই	ই	ি	i
6	ii	iː	ई	ई	ई	ઈ	ई	ஈ	ഈ	ఈ	ಈ	ঈ	ঈ	ঈ		
7	u	u, ʊ	उ	उ	उ	ઉ	उ	உ	ഉ	ఉ	ಉ	উ	উ	উ	ু	u
8	eu	ɯ	–	–	–	–	–	உ	്	–	್	–	–	–		w
9	uu	uː	ऊ	ऊ	ऊ	ઊ	ऊ	ஊ	ഊ	ఊ	ಊ	ঊ	ঊ	ঊ		
10	rq	–	ऋ,ऋ	ऋ,ऋ	ऋ	–	ऋ	–	ഋ	ఋ,ఋ	ಋ,ಋ	ঋ,ৠ	ঋ	–		
11	e	e	–	–	–	–	ए	எ	എ	ఎ	ಎ	এ	এ	এ	ে	e
12	ee	eː	ए	ए,ऐ	ए	એ	–	ஏ	ഏ	ఏ	ಏ	–	–	–		
13	ea	ɛ									ಎ					
14	ei	ɛː	ऐ	–	ऐ	ઐ	ऐ	–	–	–	–	–	–	–		
15	ai	aɪ	–	ऐ	–	–	–	ஐ	ഐ	ఐ	ಐ	–	–	ঐ		
16	oi	oj	–	–	–	–	–	–	–	–		ঐ	ঐ	ওই		
17	o	o	ओ	ओ,ऑ	ओ	ઓ	ओ	ஓ	ഒ	ఒ	ಒ	ও	ও	ও	ো	o
18	oo	oː	–	–	–	–	–	ஓ	ഓ	ఓ	ಓ	–	–	–		
19	ae	ae	–	ऍ	–	ઍ	–	–	–		–	অয্	–	–		
20	au	aʊ	–	औ	–	–	–	ஔ	ഔ	ఔ	ಔ	–	–	ঔ		
21	ou	oʊ	औ	–	औ	ઔ	औ	–	–	–	–	ঔ	ঔ	–		

Table 2 ILSL and IPA symbols for stop consonants of several Indian languages

22	**k**	**k**	क	क	क	ક	क	க	ಕ	ક	ठ	क	क	क	ক	ग	**k**
23	**kh**	**kʰ**	ख	ख	ख	ખ	ख	-	ಖ	ખ	ઠ	ख	ख	ख	খ	ঘ	
24	**g**	**g**	ग	ग	ग	ગ	ग	கv	ಗ	ગ	ઢ	ग	ग	ग	গ	ঙ	**g**
25	**gh**	**gʰ**	घ	घ	घ	ઘ	घ	-	ಘ	ઘ	ઢ	घ	घ	घ	ঘ	ঘ	
26	**ng**	**ŋ**	ङ	ङ	ङ	ઙ	ङ	ங	ಙ	ઙ	ઞ	ङ	ङ	ङ	ঙ	ঙ	**q**
27	**c**	**tʃ**	च	च	च	ચ	च	ச	ಚ	ચ	ઞ	च	-	च	চ	ছ	**c**
28	**ch**	**tʃʰ**	छ	छ	छ	છ	छ	-	ಛ	છ	ઠ	छ	-	छ	ছ	ছ	
29	**j**	**dʒ**	ज	ज	ज	જ	ज	ஜ	ಜ	જ	ઝ	ज	ज,य	ज	জ,য	জ	**j**
30	**jh**	**dʒʰ**	झ	झ	झ	ઝ	झ	-	ಝ	ઝ	ઝ	झ	झ	झ	ঝ	ঝ	
31	**nj**	**ɲ**	ञ	ञ	-	ઞ	ञ	ஞ	ಞ	ઞ	ઞ	-	ञ	ञ	ঞ	ঞ	**x**
32	**tx**	**ʈ**	ट	ट	ट	ટ	ट	ட	ಟ	ટ	ઠ	ट	ट	ट	ট	ট	
33	**txh**	**ʈʰ**	ठ	ठ	ठ	ઠ	ठ	-	ಠ	ઠ	ઠ	ठ	ठ	ठ	ঠ	ঠ	
34	**dx**	**ɖ**	ड	ड	ड	ડ	ड	டv	ಡ	ડ	ઢ	ड	ड	ड	ড	ড	
35	**dxh**	**ɖʰ**	ढ	ढ	ढ	ઢ	ढ	-	ಢ	ઢ	ઢ	ढ	ढ	ढ	ঢ	ঢ	
36	**nx**	**ɳ**	ह	ह	ह	ણ	ह	ண	ಣ	ણ	ઞ	ण	ण	ण	ণ		
37	**t**	**t̪**	त	त	त	ત	त	த	ತ	ત	ઠ	त,ॎ	त,ॎ	त,ॎ	ত	ত	**t**
38	**th**	**t̪ʰ**	थ	थ	थ	થ	थ	-	ಥ	થ	ઠ	थ	थ	थ	থ	থ	
39	**d**	**d̪**	द	द	द	દ	द	தv	ದ	દ	ઢ	द	द	द	দ	দ	**d**
40	**dh**	**d̪ʰ**	ध	ध	ध	ધ	ध	-	ಧ	ધ	ઢ	ध	ध	ध	ধ	ধ	
41	**n**	**n**	न,ऩ	न,ऩ	न	ન	न	ந,ன	ನ	ન	ઞ	न	न	न	ন	ন	**n**
42	**p**	**p**	प	प	प	પ	प	ப	ಪ	પ	ઠ	प	प	प	প	প	**p**
43	**ph**	**pʰ**	फ	फ	फ	ફ	फ	-	ಫ	ફ	ઠ	फ	फ	फ	ফ	ফ	
44	**b**	**b**	ब	ब	ब	બ	ब	பv	ಬ	બ	ઢ	ब	ब	ब	ব	ব	**b**
45	**bh**	**bʰ**	भ	भ	भ	ભ	भ	-	ಭ	ભ	ઢ	भ	भ	भ	ভ	ভ	
46	**m**	**m**	म	म	म	મ	म	ம	ಮ	મ	ઞ	म	म	म	ম	ম	**m**

Table 3 ILSL and IPA symbols for other consonants

Sr. No.	ILSL	IPA	Hindi	Marathi	Rajasthani	Gujarati	Bodo	Tamil	Malayalam	Telugu	Kannada	Bengali	Assamese	Manipuri	Odia	Dzongkha	Galo
47	y	j	य,य़	य,य़	य	ય	य	ய	യ	య	ಯ	য	য়	য়,য	ୟ		y
48	r	r	र,ऱ	र,ऱ	र	ર	र	ர	ര	ర	ರ	র	ৰ,ড়	র	ର	ར	r
49	l	l	ल	ल	ल	લ	ल	ல	ല	ల	ಲ	ল	ল	ল	ଲ	ལ	l
50	lx	ɭ	–	ळ,ऴ	ळ	ળ	–	ள	ള	ళ	ಳ	–	–	–	ଳ		
51	w	ʋ	व	व	व	વ	व	வ	വ	వ	ವ	ওয়	ৱ	ৱ	ୱ,ୱ		
52	sh	ʃ	श	श	श	શ	श	–	ശ	శ	ಶ	শ	–	শ	–		
53	sx	ʂ	ष	ष	ष	ષ	ष	ஷ	ഷ	ష	ಷ	ষ	–	ষ	–	ཤ	
54	s	s	स	स	स	સ	स	ஸ	സ	స	ಸ	স	চ, ছ	স	ସ, ଶ, ଷ	ས	s
55	h	ɦ	‿	‿	‿	હ	‿	ஹ	ഹ	హ	ಹ	হ	হ	হ	ହ		h

Hindi:

1. Word final schwa is normally deleted in Hindi. Implementation of this rule can be automated.
2. The glyphs ञ and ज्ञ are pronounced as a sequence of two phones: “g y”. This translation from glyph to the phone sequence can be done by a program.

Marathi: Marathi language uses aspirated nasals. Examples are म्ह (म्हण = say), न्ह ((न्हाणे = to take bath), ऱ्ह (ऱ्हास = decay), ल्ह (उल्हास), and व्ह (व्हाळ = will become). Marathi also has dental affricates in addition to the more common palatal affricates.

Tone languages: Several languages spoken in the northeast region of India have tonal distinctions within a spoken syllable. Numerical suffixes are used to denote tones other than a low (default) tone, Dzongkha uses a high tone, Meitei (Manipuri) uses a rising tone, Adi and Galo languages use rising and falling tones in addition to the default low tone. Mizo uses 4 tones: low, high, falling, and rising. The numerical suffixes 1, 2, 3, and 4 are used to denote a high, rising, falling, rising and falling tone, respectively. A vowel with a low or neutral or natural tone does not have any numerical suffix.

Devanagari alphabet --> ILSL

अ	आ, ा	इ, ि	ई, ी	उ, ु	ऊ, ू	ए, े	ऐ, ै	ओ, ो	औ, ौ
a	aa	i	ii	u	uu	ee	ei	o	ou

क	ख	ग	घ	ङ
k	kh	g	gh	ng
च	छ	ज	झ	ञ
c	ch	j	jh	nj
ट	ठ	ड	ढ	ण
tx	txh	dx	dxh	nx
त	थ	द	ध	न
t	th	d	dh	n
प	फ	ब	भ	म
p	ph	b	bh	m

य	र	ल	व	श	ष	स	ह
y	r	l	w	sh	sx	s	h

Fig. 1 Devanagari alphabet and the associated ILSL symbols

Tables 4 and 5 list the ILSL symbols of phones that occur in a small set of languages.

2.4 Discussion

There are many more phonetic variations that are not included in the ILSL. They should be incorporated. Of course, one can ask the following question. In this era of end-to-end systems, where text in the native script is the input to a TTS system and is the output of a ASR system, are these labels relevant? One answer is: They are useful for acoustic–phonetic analysis of spoken languages. Traditional ASR and TTS systems are still being used in low power devices without Internet connection.

Table 4 ILSL and IPA symbols for a few characters and glyphs that occur in one or two languages

Sr. No.	ILSL	IPA	Hindi	Marathi	Rajasthani	Gujarati	Bodo	Tamil	Malayalam	Telugu	Kannada	Bengali	Assamese	Manipuri	Odia	Dzongkha
72	zh	ɻ	–	–	–	–	–	ழ	ഴ	–	–	–	–	–	–	
73	nd		–	–	–	–	–	ந	–	–	–	–	–	–	–	
74	nxh	ɳʰ	–	ण्	–	–	–	–	–	–	–	–	–	–	–	
75	nh	nʰ	–	न्	–	–	–	–	–	–	–	–	–	–	–	
76	mh	mʰ	–	म्	म्	–	–	–	–	–	–	–	–	–	–	
77	rh	lʰ	–	र्	–	–	–	–	–	–	–	–	–	–	–	
78	lh	lʰ	–	ल्	–	–	–	–	–	–	–	–	–	–	–	ཧ
80	wh	wʰ	–	व्	व्	–	–	–	–	–	–	–	–	–	–	
81	cx	t͡ʃ	–	च	–	–	–	–	–	–	–	–	–	–	–	
82	jx	d͡ʃ	–	ज	–	–	–	–	–	–	–	–	–	–	–	
83	x	x	–	–	–	–	–	–	–	–	–	–	শ,ষ,স	–	–	

3 Conclusions

Indian languages use “phonetic scripts”; there is a good correspondence between the characters and speech sounds. Almost all Indian languages use scripts derived from the Brahmi script. Thus, there is great similarity in the glyphs. However, there are minor variations as well. Since the characters of the Roman alphabet are unable to label all speech sounds of Indian languages without ambiguity, a common set of labels for speech sounds of Indian languages was devised. This set, called Indian Language Speech Labels, has been extensively used by major speech groups in India, and has become a de facto standard. Such a common label set fosters borrowing of speech data from other languages while implementing systems in a low resourced language. This hopefully will accelerate the development of speech technologies in many Indian languages.

Table 5 ILSL and IPA symbols for a few consonants that occur in a smaller set of languages

Sr. No.	ILSL	IPA	Hindi	Marathi	Rajasthani	Gujarati	Bodo	Tamil	Malayalam	Telugu	Kannada	Bengali	Assamese	Manipuri	Odia	Dzongkha
56	**kq**	**q**	**क़**	**क़**	–	–	–	–	–	–	–	–	–	–	–	
57	**khq**	**x**	**ख़**	**ख़**	–	–	–	–	–	–	–	–	–	–	–	
58	**gq**	**ɣ**	**ग़**	**ग़**	–	–	–	–	–	–	–	–	–	–	–	
59	**z**	**z**	**ज़**	**ज़**	–	–	–	–	–	–	–	–	–	–	–	ཟ
60	**jhq**	**ʒ**	**झ़**	**झ़**	–	–	**झ**	–	–	–	–	–	–	–	–	
61	**dxq**	**ɽ**	**ड़**	**ड़**	**ड़**	–	**ड़**	–	**ऽ**	–	–	**ড়**	–	–	**ଡ଼**	
62	**dxhq**	**ɽʰ**	**ढ़**	**ढ़**	**ढ़**	–	**ढ़**	–	–	–	–	**ঢ়**	**ঢ়**	–	**ଢ଼**	
63	**dhq**		–	–	**ध**	–	–	–	–	–	–	–	–	–	–	
64	**f**	**f**	**फ़**	**फ़**	–	–	–	ஃப	–	–	–	**ফ়**	–	–	–	
65	**bq**		–	–	**ब**	–	–	–	–	–	–	–	–	–	–	
66	**yq**		–	–	–	–	–	–	–	–	–	**য়**	–	–	–	
67	**nq**		–	–	–	–	–	–	ന	–	–	–	–	–	–	
68	**rx**	**ɾ**	–	–	–	–	–	ற	റ	ఱ	ಱ	–	–	–	–	ར
69	**sq**		–	–	**स**	–	–	–	–	–	–	–	–	–	–	
70	Nasa Liza tion		ँ	ँ	ँ	ઁ	–	–	–	–	–	ঁ	ঁ	ঁ	ଁ	
79	**Q**		ं	ं	ं	ં	ं	–	ം	ం	ಂ	ং	ং	ং	ଂ	
71	**h**		ः	ः	ः	ઃ	ः	–	ഃ	ః	ಃ	ঃ	ঃ	–	ଃ	

Acknowledgements A work of this nature would not be possible without the contributions and support of many speech researchers in the country. I would like to thank Prof. Peri Bhashar Rao, Dr. Nandini Bondale, and Dr. Priyankoo Sarmah for their valuable inputs. The ASR and TTS projects, sponsored by TDIL, Government of India, gave a strong push to the standardization process.

References

1. International Phonetic Association: the IPA chart. https://www.internationalphoneticassociation.org/content/ipa-chart
2. Comparison of ASCII encodings of the international phonetic alphabet. Wikipedia. https://en.wikipedia.org/wiki/Comparison_of_ASCII_encodings_of_the_International_Phonetic_Alphabet
3. Klautau A (2001) ARPABET and the TIMIT alphabet. https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf
4. SAMPA. Wikipedia. https://en.wikipedia.org/wiki/SAMPA
5. Li A, Chen X, Sun G, Hua W, Yin Z, Zu Y, Zheng F, Song Z (2000) The phonetic labeling on read and spontaneous discourse corpora, pp 724–727
6. Acharya, multilingual computing for literacy and education. IIT Madras. https://www.acharya.gen.in:8080/multi_sys/exist_codes.php
7. The Graphic and Intelligence based Script Technology (GIST). https://www.cdac.in/index.aspx?id=mlc_gist_pcigcard
8. Indian script code for information interchange. https://en.wikipedia.org/wiki/Indian_Script_Code_for_Information_Interchange
9. Indian Script Code for Information Interchange (ISCII). https://varamozhi.sourceforge.net/iscii91.pdf
10. Samudravijaya K, Krishnan S, Rao PVS (1994) A general purpose phonetically labeled speech database in Hindi. In: Agrawal SS (ed) Proceedings of the international conference on applications of information technology to South Asian languages. BPB Publications, New Delhi, pp 127–134
11. Hussain T, Samudravijaya K (2011) Comparison and usefulness of ASR11 scheme over previous schemes for transliteration and label set purposes for Indian languages. Presented at the 39th All India DLA conference, Punjabi University, Patiala, 14–16th June 2011
12. Speech-based automated commodity price and weather information helpline for twelve Indian states. https://asrmandi.wixsite.com/asrmandi
13. Ramani B, Lilly Christina S, Anushiya Rachel G, Sherlin Solomi V, Nandwana MK, Prakash A, Aswin Shanmugam S, Krishnan R, Kishore SP, Samudravijaya K, Vijayalakshmi P, Nagarajan T, Murthy HA (2013) A common attribute based unified HTS framework for speech synthesis in Indian languages. In: Proceedings of SSW8: 8th ISCA speech synthesis workshop, Aug 31–Sep 2, 2013, Barcelona, Spain, pp 291–296
14. Samudravijaya K, Murthy HA (2012) Indian language speech sound label set (ILSL12), developed by Indian Language TTS Consortium & ASR Consortium. https://www.iitm.ac.in/donlab/tts/downloads/cls/cls_v2.1.6.pdf

Author Index

A
Acharya, Sathwik, 211
Agashe, Saaket, 295
Aggarwal, Rajesh Kumar, 409
Amitesh, Om, 211
Ashwini, 211

B
Banerjee, Kaushik, 55, 67
Bhat, Ashwini, 145
Bhoomika, 261
Biswas, Anupam, 109
Borgohain, Junmoni, 199, 285
Borwankar, Saumya, 37
Brazier, Charles, 157

C
Chakraborty, Samanway, 223
Chakraborty, Soubhik, 249
Chapparband, Muktharunisa, 145
Chatterjee, Subarna, 85, 119
Clarinda Devi, Thangjam, 425

D
Das, Dipankar, 223, 367, 381
Das, Himanish Shekhar, 395
Das, Rubel, 367
d"Eon, Jason, 49
Deshmukh, Saurabh, 187
Devalla, Vihar, 211
Dhall, Aditya, 235
Dhar, Ankita, 133
Dogra, Manmohan, 37
Domala, Jayashree, 37
Dumpala, Sri Harsha, 49

F
Farheen, Nida, 85, 119

G
Gangashetty, Suryakanth V., 3
Ganguly, Sandipan, 223
Gayen, Pinaki, 285
Ghosh, Dipak, 55, 67
Gor, Krutarth, 171

K
Karambelkar, Gouri, 199
Kaushik, Manoj, 99
Kim, Hanah, 341
Koolagudi, Shashidhar G., 235
Kulkarni, Megha Gururaj, 145
Kumar, Ankit, 409
Kumar, Ranjeet, 109

M
Mankad, Sapan H., 171
Modak, Akash, 223
Mondal, Souvik, 381
Mukherjee, Himadri, 133
Mullick, Raju, 199

N
Naik, Dinesh, 355

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

A. Biswas et al. (eds.), *Advances in Speech and Music Technology*, Advances in Intelligent Systems and Computing 1320, https://doi.org/10.1007/978-981-33-6881-1

Nandi, Sukumar, 133
Narasimhan, S. V., 17, 275
Nataraja, N. P., 275
Nemade, Milind U., 325
Nisha, Kavassery Venkateswaran, 261

O
Obaidullah, Sk. Md., 133
Oore, Sageev, 49

P
Patel, Anshul, 171
Patil, Shubham, 295
Patnaik, Priyadarshi, 199, 285
Patranabis, Anirban, 55, 67
Paul, Rajat, 435
Phadikar, Santanu, 133
Pulijala, Amulya Sri, 3

R
Rani, Shikha, 99
Ray, Antariksha, 381
Roy, Kaushik, 133
Roy, Pinki, 395

S
Sameeksha, D. K., 145
Samudravijaya, K., 435, 449
Sarbajit Singh, Leihaorambam, 425
Sengupta, Ranjan, 55, 67
Shah, Anuj, 171
Shah, Hitesh B., 309
Shete, Shambhavi, 187
Shukla, Pranjala, 249
Singh, Sachin Kumar, 25
Singh, Yeshwant, 109
Soumya, Mahesh, 17
Srinivasa Murthy, Y. V., 235

T
Thaoroijam, Kabita, 425

V
Varnagar, Chintan R., 309
Vijaya Krishna, A., 145
Vyas, Kartik, 355

W
Widmer, Gerhard, 157

Y
Yadav, Mainejar, 25
Yadav, Vrinda, 99