IEEE Signal Processing Magazine

[VOLUME 28 NUMBER 6 NOVEMBER 2011]

MULTIMEDIA QUALITY ASSESSMENT

A WORLD OF APPLICATIONS

GOLDEN ERA OF SIGNAL PROCESSING

TRENDS IN MULTIMEDIA AND MACHINE LEARNING SP

TELEPRESENCE: VIRTUAL REALITY IN THE REAL WORLD

SUB-NYQUIST SAMPLING

IEEE Signal Processing Society

IEEE

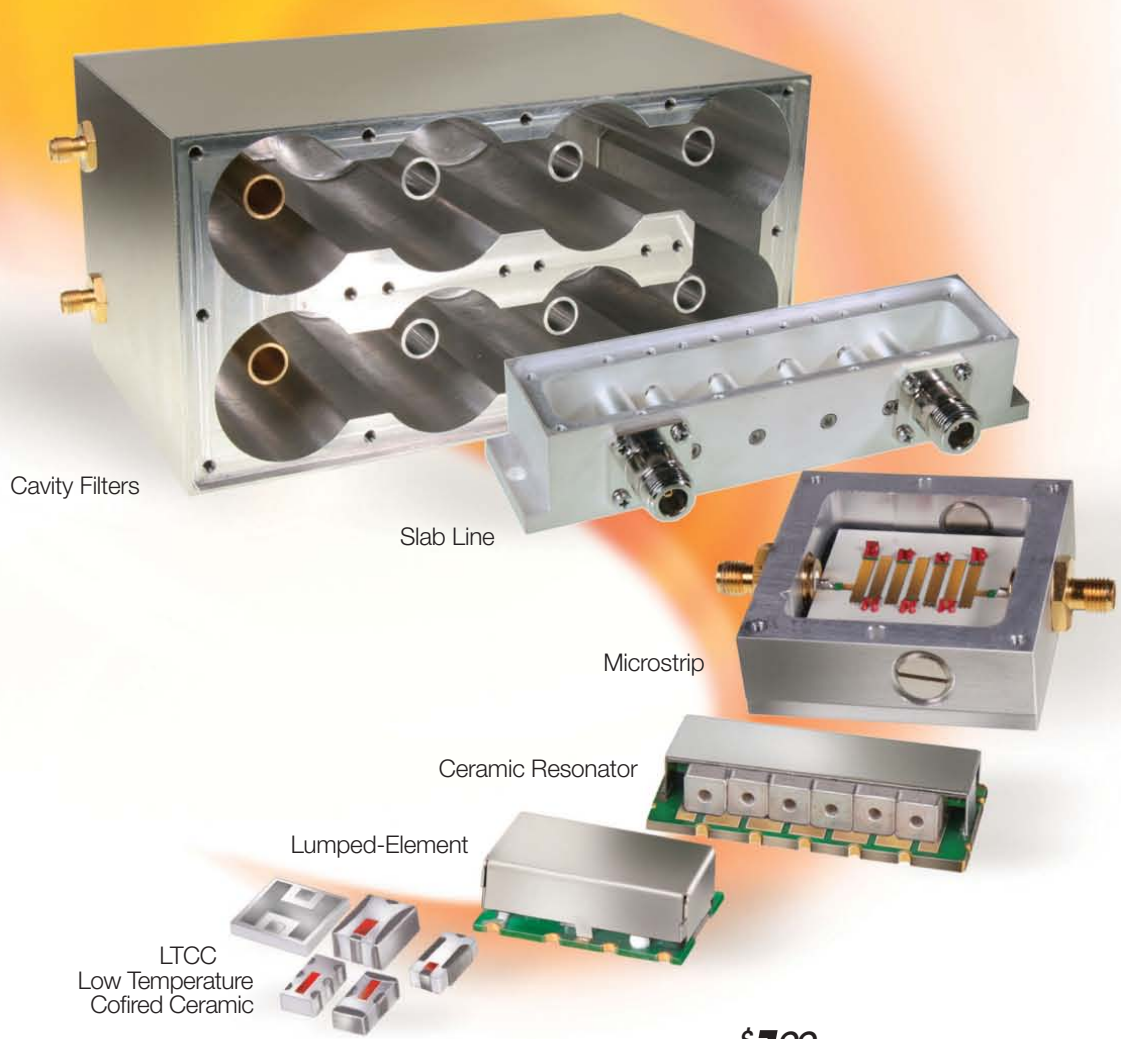
IEEE SIGNAL PROCESSING MAGAZINE

MULTIMEDIA QUALITY ASSESSMENT

VOLUME 28 NUMBER 6 NOVEMBER 2011

FILTER SOLUTIONS

DC to 15 GHz



Over 300 Models IN STOCK ...Immediate Delivery! from $1.99 ea. 10-49

Different needs demand different technologies, and the Mini-Circuits RF/microwave filter lineup delivers. Over 300 proven solutions, from DC to 15 GHz, are standing by, ready to ship. High-pass or low-pass, band-pass or band-stop, in coaxial, surface-mount, or plug-in packages. Across the board, our filters achieve low insertion loss and low VSWR in the passband and high attenuation in the rejection band. Just go to minicircuits.com for more information. If you need a specific performance and want to search our entire model database, including engineering models, click on Yoni2, our exclusive search engine.

In Yoni2, you can enter the response type, connection option, frequency, insertion loss, or any other specifications you have. If a model cannot be found, we understand the sense of urgency. So contact us, and our engineers will find a quick, cost-effective, custom solution and deliver simulation results within a few days.

Yoni2® U.S. Patent 7739260, 7761442 ***The Design Engineers Search Engine...*** *finds the model you need, Instantly.*

Mini-Circuits…we're redefining what VALUE is all about!

Mini-Circuits®
ISO 9001 ISO 14001 AS9100

P.O. Box 350166, Brooklyn, New York 11235-0003 (718) 934-4500 Fax (718) 332-4661

Yoni2® U.S. Patents 7739260, 7761442 ***The Design Engineers Search Engine*** *finds the model you need, Instantly* • For detailed performance specs & shopping online see minicircuits.com

IF RF MICROWAVE COMPONENTS

484 Rev B

VOLUME 28 NUMBER 6

CONTENTS

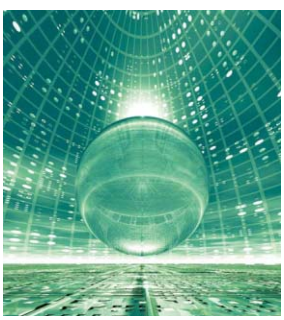
SPECIAL SECTION—MULTIMEDIA QUALITY ASSESSMENT

FEATURE

COLUMNS

DEPARTMENTS

COVER ©GETTY IMAGES



SCOPE: *IEEE Signal Processing Magazine* publishes tutorial-style articles on signal processing research and applications, as well as columns and forums on issues of interest. Its coverage ranges from fundamental principles to practical implementation, reflecting the multidimensional facets of interests and concerns of the community. Its mission is to bring up-to-date, emerging and active technical developments, issues, and events to the research, educational, and professional communities. It is also the main Society communication platform addressing important issues concerning all members.

IEEE SIGNAL PROCESSING MAGAZINE (ISSN 1053-5888) (ISPREG) is published bimonthly by the Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, 17th Floor, New York, NY 10016-5997 USA (+1 212 419 7900). Responsibility for the contents rests upon the authors and not the IEEE, the Society, or its members. Annual member subscriptions included in Society fee. Nonmember subscriptions available upon request. Individual copies: IEEE Members $20.00 (first copy only), nonmembers $141.00 per copy. Periodicals postage paid at New York, NY, and at additional mailing offices. Postmaster: Send address changes to IEEE Signal Processing Magazine, IEEE, 445 Hoes Lane, Piscataway, NJ 08854 USA. Canadian GST #125634188

Printed in the U.S.A.

Digital Object Identifier 10.1109/MSP.2011.941849

[from the EDITOR]

Li Deng
Editor-in-Chief
deng@microsoft.com
http://signalprocessingsociety.org/publications/periodicals/spm

Shining Bright: The Golden Era of Signal Processing

Three years ago, I wrote my inaugural editorial, "Embracing a New Golden Age of Signal Processing," [1] for our *IEEE Signal Processing Magazine (SPM)*. Today, while writing this farewell editorial and in representing our entire *SPM* editorial team, I can proudly say the golden era has not only arrived for us to embrace and celebrate, but it is also shining bright and is here to stay.

My service for *SPM* started in 2004, when Prof. Ray Liu, then editor-in-chief, invited me to be the lead guest editor for a special issue. Since then, *SPM* has been a main focus of my service to the IEEE Signal Processing Society (SPS) and to the SP community. When working as an editorial board member and area editor under the leadership of my predecessor, Prof. Shih-Fu Chang, I witnessed the immeasurable vibrancy, invigorating energy, and unbounded intellectual landscape of our SP community. During 2007–2008, with Prof. Chang's guidance, I initiated the effort in expanding the scope and technical fields of SP [2]. This led to substantial broadening of the article coverage in *SPM* along the two axes of "signal" and "processing" [3]. In the meantime, while helping Prof. Chang to solicit potential articles for *SPM*, I interacted with several pioneers in various technical areas pertinent to SP. These interactions provided me with the opportunity to learn, analyze, and appreciate a wide range of SP-enabled future wants and needs (e.g., [4]). In my own work environment within a major computer software company, SP methods and applications as defined in the expanded scope had also permeated every corner. Our community had clearly come to realize that while SP played an integral part in the technological development of television, telephone, communication, multimedia, space travel, and computers, more exciting challenges and opportunities would lie ahead for SP in broad areas such as intelligent communication; natural human-machine interface; universal language translation; biomolecular information processing; automated navigation; efficient generation/distribution/consumption of "green" energy; intelligent sensor and human

Digital Object Identifier 10.1109/MSP.2011.942544
Date of publication: 1 November 2011

Digital Object Identifier 10.1109/MSP.2011.942735

networks; global financial market analysis; and much more [4], [5]. All these, together with the numerous SP-empowered technological advancements—e.g., mobile devices becoming ubiquitous, multicore and cloud computing going mainstream, Web search turning intelligent—already brewing in the midst of economic recession three years ago, heralded a new tech boom and a bright era ahead in our field of SP.

Indeed, during the past three years, we have witnessed tremendous growth in signal processing at the global scale. I was honored to lead an energetic, diligent, creative, and productive editorial team to embrace the vitality of our SP community in this golden age. The magazine served not only as an educational tool but also as a catalyst in advancing SP technology. Our articles exemplified and embodied technical rigor and new trends of SP as well as the extraordinary variety of SP applications in our daily lives and their societal impact.

Our editorial team took a unique approach to running *SPM*. We took risks, pushed the limit, and we were eager to innovate and try things that had never been done before. We embraced the motto that it is more fun being movers and shakers than being followers and being incremental. We held the attitude that if we fail, let it be, but if we succeed, we would win big. (Don't we all run research groups and do SP research in the same way?) One significant innovation we engendered over the past two years is the translation editions of *SPM* into Chinese and Brazilian Portuguese, where we saw huge emerging SP engineering bases and potential explosive readership growth. This is the first time in history that IEEE publications have done a Chinese translation. References [6]–[10] are just a few examples of a larger pool of the articles we have translated to Chinese and Brazilian Portuguese. We listened closely to our readers' feedback in Asian-Pacific countries and published side-by-side English-Chinese translation. This enables them not only to read the technical content more efficiently in their native language, but more importantly, to write better SP articles in English. To accommodate the difference between simplified and traditional Chinese styles that are both popular in Asian-Pacific countries, we created a three-column glossary for technical terms of SP in English and in the corresponding simplified and traditional Chinese pairs. As an example of the impact of this "side" glossary project within our much larger translation project, we are looking at extending our approach to all IEEE-relevant technical terms that are far beyond the scope of SP.

Another important innovation we have created and pushed hard is the use of Tag for direct and convenient access to multimedia supplementary material via smartphones, which can go with the readers everywhere. It opens a new way of linking and integrating the printed material with the author-created online content. It also opens a new opportunity for creative design of the online supplementary material (e.g., animated figures that would drastically enhance the current static figures in print and the "just-in-time" contextual appendix, references, or video/audio/handwriting tutorials, etc.). Other notable innovations we have instituted include the (ongoing) cross-Society collaboration to attract wider audiences, digital delivery of our articles, special issues focusing on emerging SP applications [5], and publications of unique types of articles. Examples of the latter are the articles reporting vastly visible SP applications (e.g., [11] and [12]), highlighting research directions (e.g., [13]–[15]), analyzing technical trends and their future (e.g., [4] and [15]), and focusing on SP education or history with a lecture-note style (e.g., [16]–[20]).

Our approach turned out to be quite successful. In the latest IEEE Annual Report, our Chinese translated edition of *SPM* is prominently featured with the following strong endorsement:

> … The first IEEE publication in Chinese, this special issue of *IEEE Signal Processing Magazine* was distributed in 2010 at the Society's International Conference on Image Processing in Hong Kong. … This Chinese translation is the first step in the Society's efforts to enhance its visibility among non-English speaking audiences. …*SPM* ranks highest among all electrical and electronic engineering journals.

In the summers of 2010 and 2011, *SPM*'s accomplishments were reflected in Thomson Reuters' *Journal Citations Reports (JCR)*, generally accepted as the world's most influential source of information about peer-reviewed publications. *SPM*'s impact factor has dramatically increased from 3.76 in 2008 (ranked ninth; see the comparison group below) to 4.91 in 2009 (moved to first place) and further to 5.86 in 2010 (continues to rank first). Comprehensive compilation in the *JCR* results shows that our *SPM*'s top rank is among all publications in the broad electrical and electronics engineering field (247 of them in total, including 147 of IEEE's) the two most recent years in a row. The five-year long-term impact factor of *SPM* also jumped from 5.95 (in 2009) to 6.89 (in 2010) as a result of highly cited, most recent *SPM* articles. Further, the Article Influence Score of *SPM* continues to rank number one among the 247 journals, again two most recent years in a row—2.48 in 2009, jumping to 3.18 in 2010. This is an outstanding record and an honor brought to our Society and community, which in turn validates our novel approach to running *SPM*.

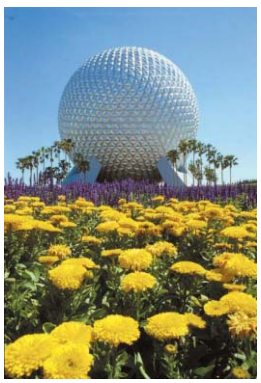
Behind the success and top-ranking honor are the real heros: our editorial team members. I thank them for taking the journey with me in our relentless pursuit of excellence and in pushing the boundaries of innovation in running our *SPM*. Area Editors Dan Schonfeld (special issues), Antonio Ortega (feature articles), Ghassan AlRegib (column/forum), and Jane Wang (together with Min Wu, as e-newsletter area editors) deserve special recognition and appreciation. Their dedicated service, infectious enthusiasm, and selfless sacrifice over the past three years have made our *SPM* what it is today. They, together with their associate editor teams, have been tirelessly working with me towards the common and clear goal of making *SPM* the best among the best. Thanks also go to our *SPM* editorial board for their guidance in reviewing white papers, providing feedback, and

2012 IEEE International Conference on Image Processing

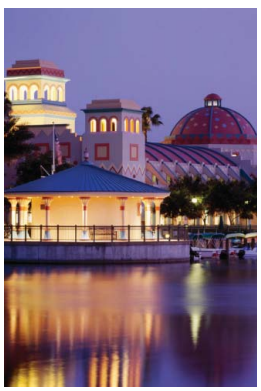
Disney's Coronado Springs Resort, Lake Buena Vista, Florida, U.S.A.

September 30-October 3, 2012

© Disney



© Disney



© Disney

© Disney

IEEE

IEEE Signal Processing Society®

CALL FOR PAPERS

The International Conference on Image Processing (ICIP), sponsored by the IEEE Signal Processing Society, is the premier forum for the presentation of technological advances and research results in the fields of theoretical, experimental, and applied image and video processing. ICIP-2012, the nineteenth in the series that has been held annually since 1994, will bring together leading engineers and scientists in image and video processing from around the world. Research frontiers in fields ranging from traditional image processing applications to evolving multimedia and video technologies are regularly advanced by results first reported in ICIP technical sessions. Topics include, but are not limited to:

- **Image/video coding and transmission:** Still image and video coding, stereoscopic and 3-D coding, distributed source coding, source/channel coding, image/video transmission over wireless networks
- **Image/video processing:** Image and video filtering, restoration and enhancement, image segmentation, video segmentation and tracking, morphological processing, stereoscopic and 3-D processing, feature extraction and analysis, interpolation and super-resolution, motion detection and estimation, color and multispectral processing, biometrics
- **Image formation:** Biomedical imaging, remote sensing, geophysical and seismic imaging, optimal imaging, synthetic-natural hybrid image systems
- **Image scanning, display, and printing:** Scanning and sampling, quantization and halftoning, color reproduction, image representation and rendering, display and printing systems, image quality assessment
- **Image/video storage, retrieval, and authentication:** Image and video databases, image and video search and retrieval, multimodality image/video indexing and retrieval, authentication and watermarking
- **Applications:** Biomedical sciences, mobile imaging, geosciences and remote sensing, astronomy and space exploration, document image processing and analysis, other applications

Paper Submission: Prospective authors are invited to submit papers of not more than four (4) pages including results, figures and references. Papers will be accepted only by electronic submission at **www.icip2012.com**.

Submission of papers: **January 12, 2012**
Notification of acceptance: **April 13, 2012**
Submission of camera-ready papers: **May 18, 2012**

Tutorials: Tutorials will be held on September, 30, 2012. Brief proposals should be submitted by January 9, 2012 to Prof. Lina Karam and Prof. Andreas Savakis at the conference web site. Proposals for tutorials must include a title, an outline of the tutorial and its motivation, a short description of the material to be covered, contact information including name, affiliation, email, and mailing address for each presenter, and a two-page CV for each presenter.

Special Sessions: Special sessions proposals should be submitted by December 2, 2011, to Prof. David Taubman at the conference web site. Proposals for special sessions must include a topical title, rationale, session outline, contact information for the session chair(s) and authors who have agreed to present a paper in the session, and a tentative title and abstract of each paper.

Special sessions proposals due: **December 2, 2011**
Notification of special sessions acceptance: **December 23, 2011**
Tutorial proposal due: **January 9, 2012**
Notification of tutorial acceptance: **February 13, 2012**

www.icip2012.com

General Chair
Prof. Eli Saber
Rochester Institute of Technology

Technical Program Chairs
Prof. Sheila Hemami
Cornell University
Prof. Gaurav Sharma
University of Rochester

Finance Chair
Prof. Sohail Dianat
Rochester Institute of Technology

Plenary Chair
Prof. Dan Schonfeld
University of Illinois at Chicago

Tutorials Chairs
Prof. Lina Karam
Arizona State University
Prof. Andreas Savakis
Rochester Institute of Technology

Special Sessions Chair
Prof. David Taubman
University of South Wales

Publications Chairs
Dr. Raghuveer Rao
Army Research Labs
Dr. Prudhvi Gurram
MBO Partners

Publicity Chairs
Prof. Keigo Hirakawa
University of Dayton
Dr. Farhan Baqai
Apple Inc.

Industrial Program Chairs
Dr. Khaled El-Maleh
Qualcomm Inc.
Dr. Amir Said
Hewlett-Packard Laboratories

Exhibits Chair
Dr. Amit Singhal
Eastman Kodak
Dr. Yaowu Xu
Google

Awards Chair
Dr. John Apostolopoulos
Hewlett-Packard Laboratories

Local Arrangements Chair
Prof. Hassan Foroosh
University of Central Florida

International Liaisons
Prof. Murat Tekalp
Koc University
Prof. Alex Kot
Nanyang Technological University
Prof. Ricardo de Queiroz
Universidade de Brasilia

IEEE Orlando
Prof. Xun Gong
University of Central Florida

Conference Management
Ms. Billene Mercer
Conference Management Services, Inc.

setting directions. All of our editors and board members have been involved in identifying hot topics, recruiting the best authors to write papers on these hot topics, coordination of paper writing and reviewing, guarding the paper acceptance threshold, and making suggestions on how to innovate. And, of course, the fundamental credit should go to the authors, whose high-quality papers made the impact possible, and to the guest editors, reviewers, and readers. Some authors not only wrote the original papers in English but also sacrificed their time helping with proofreading (an extremely crucial step) of the translated versions.

Here I wish to express my special, wholehearted thanks to Linda Cherry, in her role of SPS publication manager, as well as to her staff for the tremendous contributions, especially those to the *SPM* translation work. Linda also spear-headed the efforts to implement more cutting-edge design of our *SPM* covers, on digital delivery of our articles, and on the industry-friendly articles that are extremely popular. Her enthusiasm, creativity, dedication, and work ethic are just admirable. I counted over 2,200 e-mails between us during the past two years within my e-mail folder just on our Chinese translation project alone. Senior Managing Editor Geri Krolin-Taylor, Managing Editor Jessica Barragué, Art Director Janet Dudar, Production Coordinator Terry Smith, and Supervisor of Production/Periodicals Louis Vacca once again demonstrated a high level of professionalism in processing our highly dynamic technical content and in bringing it to the readers in the best possible way no matter how much time or effort it took.

I am also greatly indebted to Prof. Shih-Fu Chang and Prof. Ray Liu, my first- and second-level predecessors, respectively, who advised and helped me not only during the editor-in-chief transition period but also throughout my term whenever I needed them. In 2003, Ray historically revolutionized *SPM* to its current structure as then editor-in-chief and brought it to the number one rank. Since then, *SPM* has been on the top-ranking list, thanks to Shih-Fu and his team's continued efforts to make it better. Several of my predecessors passed *SPM* to me already in excellent shape—it ranked 33rd, 16th, second, first, third, tenth, 12th, fifth, and ninth among over 200 journals in the field worldwide from the years 2000–2008. Our current *SPM* team built upon these earlier successes, not just to continue the tradition but also to aim at an even higher goal and to reach the goal with a new wave of innovations. Further, I thank Mos Kaveh, in his role as SPS president, for his support and advice as well as for taking me with him on the memorable tour of China in fall 2010 to promote SP and *SPM's* first Chinese edition. Finally, my appreciation goes to Prof. Ali Sayed, in his role of SPS vice president, publications and SPS Executive Director Mercy Kowalczyk, for their advice on the established rules and procedures.

After this issue, *SPM* will be in the capable hands of our new editorial team. Let me take this opportunity to welcome Prof. Abdelhak Zoubir, who will soon serve as the new editor-in-chief. He will be joined by new area editors and associate editors whom he will announce soon. Given that the technologies we created and continue to create are changing the world at a pace never matched in history, there is no doubt in my mind that the new team's collective wisdom, vision, and leadership will continue our *SPM*'s current successful momentum, ensure that the golden age of our field and *SPM* is long lasting, and move *SPM* to an even higher level with more innovations to come. Indeed, the opportunity—ubiquity of SP in our modern information age—welcomed the advent of the golden era, but innovations drive it to sustain and to shine brighter.

Since I started my voluntary work for *SPM* seven years ago, my service to the SP community has never been more exciting. The past three short years have been truly exhilarating, but the term for me and my team is over, and *SPM* needs new blood. Still with full energy and the desire to continue serving our Society, I thank our SPS Publications Board for appointing me as the new editor-in-chief of *IEEE Transactions on Audio, Speech, and Language Processing (T-ASLP)*. This allows me to return to my home research area, continuing my service to *T-ASLP*, which was put on hold for several years due to my work for *SPM*. This is a new prime opportunity to explore different kinds of innovations in a new capacity and in a new era of the even faster pacing technological age. So we will still be in touch, especially for those *SPM* audiences who also read *T-ASLP*. Thank you all for sharing the exciting past three years with me and thank you also in advance for supporting the new *SPM* management team.

REFERENCES

[1] L. Deng, "Embracing a new golden age of signal processing," *IEEE Signal Processing Mag.,* vol. 26, no. 1, Jan. 2009.

[2] L. Deng, "Expanding the scope of signal processing," *IEEE Signal Processing Mag.,* vol. 25, no. 3, May 2008.

[3] L. Deng, "Cross-pollination in signal processing technical areas," *IEEE Signal Processing Mag.,* vol. 26, no. 6, Nov. 2009.

[4] J. Treichler, "Signal processing: A view of the future, Parts I and II," *IEEE Signal Processing Mag.,* vol. 26, no. 2, Mar./May 2009.

[5] D. Schonfeld, "The evolution of signal processing," *IEEE Signal Processing Mag.,* vol. 27, no. 5, Sept. 2010.

[6] E. J. Candès et al., "An introduction to compressive sampling," *IEEE Signal Processing Mag.,* vol. 25, no. 2, Mar. 2008.

[7] M. Zibulevsky et al., "L1–L2 optimization in signal and image processing," *IEEE Signal Processing Mag.,* vol. 27, no. 3, May 2010.

[8] M. N. Wernick et al., "Machine learning in medical imaging," *IEEE Signal Processing Mag.,* vol. 27, no. 4, July 2010.

[9] X. He et al., "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Mag.,* vol. 25, no. 5, Sept. 2008.

[10] D. Blei et al., "Probabilistic topic models," *IEEE Signal Processing Mag.,* vol. 27, no. 6, 2010.

[11] R. Schneiderman, "DSPs are helping to make it hard to get lost," *IEEE Signal Processing Mag.,* vol. 26, no. 6, Nov. 2009.

[12] R. Schneiderman, "SETI—Are we (still) alone?" *IEEE Signal Processing Mag.,* vol. 27, no. 2, Mar. 2010.

[13] J. Baker et al., "Research developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Processing Mag.,* vol. 26, no. 3, May 2009.

[14] J. Baker et al., "Updated MINDS report on speech recognition and understanding, Part 2," *IEEE Signal Processing Mag.,* vol. 26, no. 4, July 2009.

[15] A.-J. van der Veen et al., "Trends' expert overview sessions revived at ICASSP 2011," *IEEE Signal Processing Mag.,* vol. 28, no. 5, Sept. 2011.

[16] P. Hart, "How the Hough transform was invented," *IEEE Signal Processing Mag.,* vol. 26, no. 6, Nov. 2009.

[17] D. Yu et al., "Solving nonlinear estimation problems using splines," *IEEE Signal Processing Mag.,* vol. 26, no. 4, Nov. 2009.

[18] L. Xiao et al., "A geometric perspective of large-margin training of Gaussian models," *IEEE Signal Processing Mag.,* vol. 27, no. 6, Oct. 2010.

[19] P. Prandoni et al., "From Lagrange to Shannon … and back: Another look at sampling," *IEEE Signal Processing Mag.,* vol. 26, no. 5, Sept. 2009.

[20] J. Allen et al., "Speech perception and cochlear signal processing," *IEEE Signal Processing Mag.,* vol. 26, no. 4, July 2009. SP

[president's MESSAGE]

Mostafa (Mos) Kaveh
2010–2011 SPS President
mos@umn.edu

Increasing the Visibility of Signal Processing

The IEEE is well into its five-year Public Visibility Initiative that was championed by 2007 IEEE President, and former Signal Processing Society President, Leah Jamieson. A key aim of the initiative has been to increase the public's awareness and understanding of the contributions of the engineering and computing disciplines and professions for the benefit of humanity. Signal processing has certainly done its fair share to make possible the way we live, communicate, and play. A few months ago, I had the opportunity to discuss some of these contributions with the IEEE Public Visibility staff and consultants. Our conversation was both challenging and energizing. I expect that any reader from our community will sympathize with the challenge of explaining to the general public, even those who are technically savvy, the increasingly expanded elements of the signal processing field. As usual, this challenged feeling quickly morphed into energy and excitement once I settled into describing examples of the vast array of applications and products that have been made possible by advances in signal processing. Collectively, we who are IEEE Signal Processing Society members and signal processing professionals can help increase the visibility of this vibrant technology by communicating to students and to the public at large, the central enabling role it plays in our technology-capable, everyday lives.

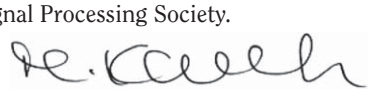
The IEEE Signal Processing Society has had, on occasion, the opportunity to recognize and publicize an historical event made possible by contributions from the signal and speech processing communities. Last summer, the IEEE Board of Directors approved the IEEE History Committee's recommendation for an IEEE Milestone in Electrical Engineering and Computing by commemorating the first real-time speech communication over a packet-switched network. The IEEE Signal Processing Society, following the recommendation of the Speech and Language Processing Technical Committee, together with the IEEE Boston Section, are cosponsors of this milestone as represented by a plaque at the MIT Lincoln Laboratory, the site of the first such transmission to the USC Information Sciences Institute in 1974. Congratulations to Cliff Weinstein and colleagues for this successful milestone nomination.

The above milestone sponsorship was one example of collaboration between IEEE's Technical Activities (TA), represented by the Signal Processing Society, and the Member and Geographical Activities (MGA) represented by the Boston Section. A major collaboration between these two pillars of the IEEE took place in August at the 2011 IEEE Sections Conference. For the first time, this triennial MGA event had significant presence from the IEEE Societies and technical councils. It was an excellent opportunity to explore and share best practices for membership development and services that are responsive to what most members demand—technical, professional, and geographical value propositions.

Congratulations and thanks to General Chair Benoit Macq and the organizers of ICIP2011 and to the accompanying THEMES on emerging technologies for video compression for their outstanding technical and social programs. The Society's Board of Governors (BoG) also had its fall meeting at ICIP in Brussels. The Board certified the elections of its three new members-at-large with 2012–2014 terms of service. The Board also discussed the possibility of initiating an organizational review of the Society. The IEEE Signal Processing Society, particularly over the past two decades, has grown organically in products, volunteer activities, and employed staff. A holistic review of the way the Society is structured and operates will appear to be beneficial at a time of changing demographics and business models for the delivery of many of the Society's products and services.

The BoG also passed a motion mourning the August passing of its former member-at-large, Alex Gershman. Alex was a brilliant scientist, a kind and generous colleague, a mentor to many, and a dedicated Society volunteer. In addition to his service as an elected member of the BoG, Alex's volunteer contributions included service as the editor-in-chief of *IEEE Signal Processing Letters*, chair of the Sensor Array and Multichannel Technical Committee, and organizer of several workshops, to name a few. He will be sorely missed, but his contributions will be lasting.

As always, I welcome your comments and suggestions on improving the products and services of your technical and professional home, the IEEE Signal Processing Society.



[SP]

Digital Object Identifier 10.1109/MSP.2011.942870
Date of publication: 1 November 2011

[special REPORTS]

John Edwards

Telepresence: Virtual Reality in the Real World

Imagine a teleconferencing system that's so realistic, so natural, that you feel you could reach out and shake a participant's hand, even though the individual may be sitting in a room hundreds or even thousands of miles away. That's the idea behind telepresence, an emerging teleconferencing technology that's designed to not only connect people together but to make them feel like they're all collaborating together inside the same room.

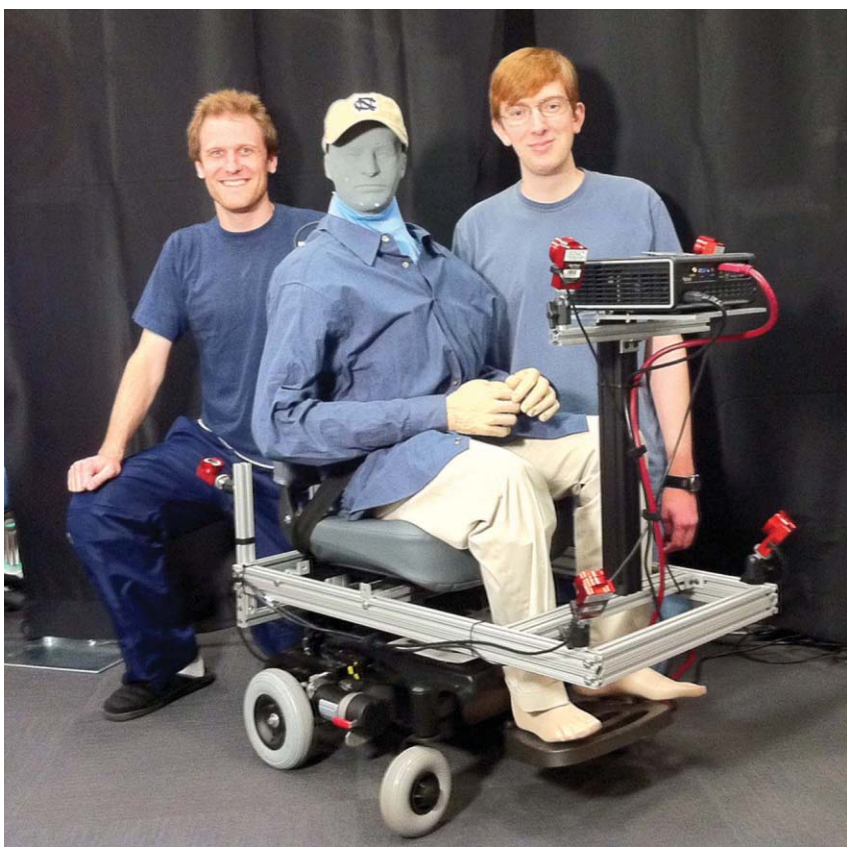
Telepresence provides an immersive meeting experience based on the highest possible levels of audio and video clarity. Ideally, participants are life-sized. Ideally, every sound, gesture, and facial expression are replicated in high-definition (HD) video and high-fidelity surround sound. The ultimate goal is to make telepresence and in-person meetings virtually indistinguishable from each other. "Telepresence has a great potential for getting people to work together collaboratively, efficiently, and naturally regardless of their physical location," observes Zhengyou Zhang, a principal researcher focusing on telepresence technologies at Microsoft Research in Redmond, Washington.

With business becoming increasingly global, and travel getting ever more expensive and cumbersome, telepresence's popularity is soaring. According to a December 2010 study by Wintergreen Research of Lexington, Massachusetts, worldwide telepresence sales are projected to reach US$6.7 billion by 2016. Yet the technology offers an important benefit that goes far beyond eliminating expensive and inconvenient travel itineraries—the ability to reach out to anyone, anywhere as if really there. "Our mind often has sparks, and those ideas need to be immediately debated and further brainstormed with colleagues and friends; otherwise, they will be just gone forever," Zhang observes.

Henry Fuchs of the University of North Carolina.

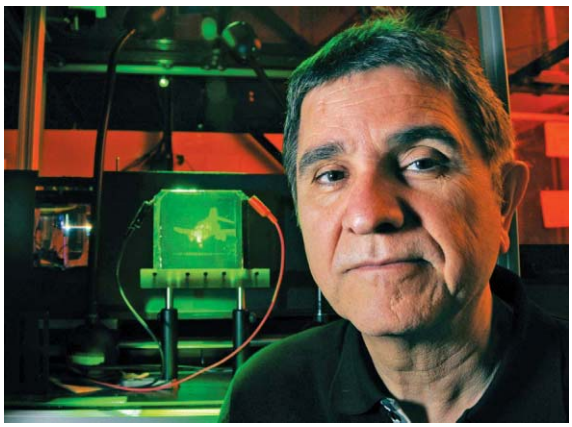
READY FOR THE REAL WORLD

Once the fodder of science fiction writers and filmmakers, telepresence is already being adopted by enterprises in sectors ranging from government to retail to manufacturing to finance to utilities. Currently available, telepresence systems are usually designed to join an enterprise's existing communications portfolio—including Web conferencing, Internet telephony, instant messaging, and social media communication tools—to help employees boost performance, increase sales, protect investments, and create rich collaborative



Ryan Schubert and Peter Lincoln (right) are grad students working on the University of North Carolina mobile avatar project.

Digital Object Identifier 10.1109/MSP.2011.941853
Date of publication: 1 November 2011



Nasser Peyghambarian, a professor of optical sciences at the University of Arizona in Tucson is shown with the hologram.

experiences with colleagues and customers. Communications integration means that participants can work face to face in a virtual meeting room while also inviting in colleagues who are limited to phone- or Web-based conferencing systems.

Telepresence systems, including cameras, displays, speakers, and other hardware and software components, are currently available from several vendors, including Cisco Systems, Tandberg, and Polycom. Prices can run anywhere from several thousand dollars to upwards of a quarter of a million dollars. Basic components in a high-end system typically include at least three HD 65-in or larger displays, several HD cameras, multiple microphones and speakers, special lighting arrays, two or more 1-Gb Ethernet ports in the room, and approximately 2–3 Mb/s of bandwidth per active screen.

While telepresence is already meeting the needs of a rapidly growing number of organizations, researchers worldwide are developing new technologies with the aim of making electronic meetings even more like "being there." Among the innovations being investigated are robot avatars that will allow people to create a virtual physical presence almost anywhere. Another important research area is three-dimensional (3-D) technology, which promises to allow colleagues, friends, and family members to pop an electronic representation of themselves into places as easily as dialing a phone. Researchers are also working on transporting realistic representations of entire meeting rooms and offices to distant locations. Zhang says telepresence's potential uses are far-reaching and limited only people's imagination. "Applications include effective team collaboration, immersive entertainment experiences, and improved customer satisfaction in marketing," he says.

ANYONE FOR ANYBOTS?

While most telepresence technologies are developed for incorporation into dedicated conference rooms, Mountain View, California-based Anybots has created a system that's designed to bring telepresence sessions directly to people at their desks or in virtually any other indoor location. The company's mobile Anybots robot is designed to function as a seeing and hearing personal avatar that can freely roll around an office, production site, sales floor, convention center, or lobby. The system is operated via a Web browser using a wireless fidelity, third generation, or fourth generation wireless connection. A video camera and speaker mounted inside the robot's head allows operators to converse with remote colleagues from a remote location while also being able to hear and see them.

An Anybot telepresence robot lets users project their presence to a work site.

Trevor Blackwell, Anybot's chief executive officer, says that ongoing advancements in video, wireless broadband, and robotics technology are making it easier and more practical for people to project their presence to remote sites. "It's now becoming possible for computers to control robot bodies in much more interesting ways than in the 1980s or 1990s," he says. "So it was time to try to build a general-purpose robot that would be useful in the office."

Blackwell states that Anybots can be used to support a variety of training, sales and customer service activities. He believes that the technology is particularly well suited for technology support services, such as helping an employee master a new software application or fix a minor computer glitch.

"Most small offices these days need a tech support person maybe one or two hours a week," he explains. He notes that small organizations often can't afford the expense of hiring a full-time support professional. Meanwhile, a part-time hire may not be available on site when needed. Blackwell says that using a telepresence robot allows a business to combine the cost and availability benefits of phone support with the interactivity attributes of a personal visit. "Being able to support people through a robot means that instead of being on the phone with someone, you can send your robot to stand behind the person, look at what they're doing on their computer, talk to them about it, look at where the wires are going, and do whatever else is necessary to troubleshoot the problem," he says.

The technology can also be put to use as an office greeter, allowing a human receptionist to remotely welcome guests without leaving the front desk vacant. "Around our office, our robot greets visitors at the door and shows them where the diet Cokes are and leads them to a seat in the conference room,"

Blackwell says. "And then it goes and gets me."

Blackwell notes that besides enabling efficient video and audio streaming, signal processing played an important role in developing the robot's physical attributes. "Our focus on signal processing has been mainly in the mobility and movement of the robot," he says. "The new technology is all about controlling the way the robot moves, making sure it doesn't crash into things, making it move smoothly and as quietly as possible, and making it seem very responsive to the remote driver even though there is some small delay over the network as they're driving it."

Such attention is vital since physical movements that are trivial for most people can be major challenges for a robot. "There are all kinds of things that can happen as you're rolling around an office—like one wheel hits a bump and then all of a sudden it's off the ground," Blackwell says. "Then, if you try to control the robot by spinning that wheel, it's going to do the wrong thing." Signal processing algorithms help the robot know what to do in specific situations.

Blackwell says that he and his team gain inspiration by observing the real world. "We spend a lot of time looking at what people do in offices and how much of that could be done remotely," he says. "We're pretty optimistic that we can replace a significant fraction of office jobs with people controlling a robot from home."

JUST LIKE BEING THERE

Robot avatar technology also figures into a telepresence system currently being developed by the BeingThere Centre, a joint project of the University of North Carolina (UNC) in Chapel Hill, North Carolina, Nanyang Technological University (NTU) in Singapore, and the Swiss Federal Institute of Technology (ETH) in Zurich. The project unites 32 leading telepresence research leaders working across three continents

A concept developed by Gregory Welch, a UNC research professor, is to create a technology that gives people the option of sending a mobile mannequin to represent themselves at a conference table. Such a unit would act as an avatar that could freely navigate to a distant environment and take on the appearance and gestures of its far-away human host. The project's ultimate goal is creating an autonomous virtual human with memory and awareness capabilities that can take the place of its host whenever he or she is absent.

An Anybot telepresence robot allows remote workers to collaborate with on-site personnel directly and informally.

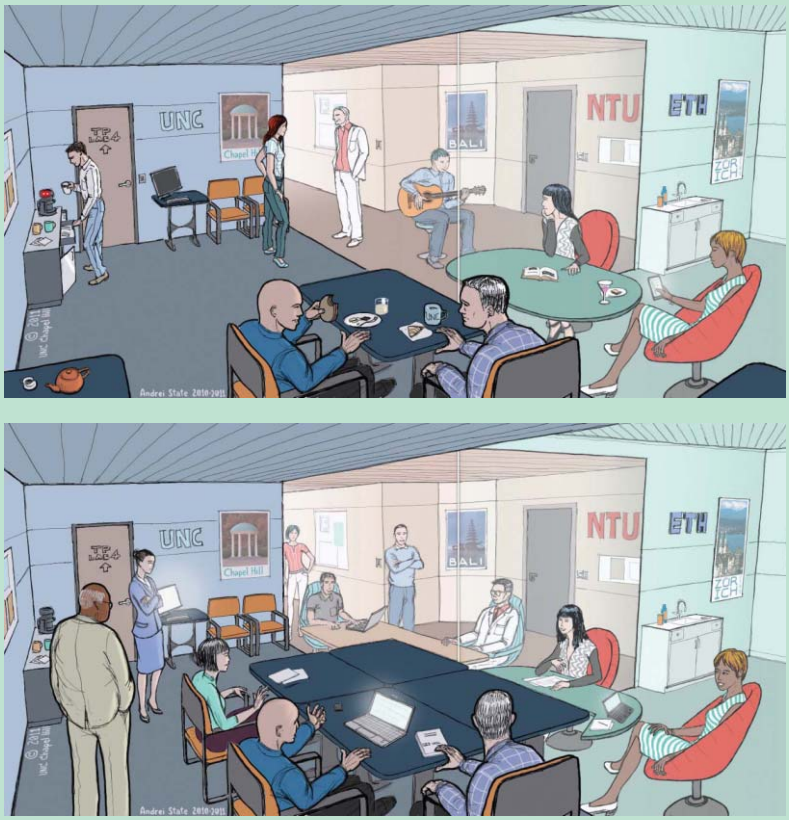
Another mobile-oriented telepresence application being developed by BeingThere Centre researchers is a portable display that can be used to bring a 3-D graphical representation of a friend, relative, or colleague to a meeting room, office, home, or other location. Like the mobile mannequin, the semitransparent display is designed to allow a user to create a virtual physical presence in a place perhaps hundreds or even thousands of miles away.

Yet another telepresence model being investigated by BeingThere Centre researchers is a conference room that connects with other similarly equipped rooms to form an entire virtual office suite. Each facility would feature wall-sized displays designed to create the illusion that the rooms are adjacent to each other and separated only by glass walls, even though they may be located in different countries. Such "glass wall" displays will provide each person in the room with a correct, personalized stereo view into the remote rooms, giving the illusion that all the of participants, both local and distant, share one common space. The idea is to allow enterprises with offices located around the world to function as if everyone were located inside the same building.

"Teleconferencing is not taking a camera image and sending it through a computer network and putting it onto a TV screen," says Henry Fuchs, a UNC computer science professor and an adjunct professor of biomedical engineering. Fuchs, who is also one of the BeingThere Centre's three codirectors (along with Prof. Markus Gross of ETH and Prof. Nadia Thalmann of NTU), feels that to provide a truly realistic immersive experience, cameras

Zhengyou Zhang with Microsoft.

John Apostolopoulos with HP.



At the BeingThere Centre, researchers are looking into the possibility of using telepresence to link geographically isolated offices into collaborative virtual suites. (Images used with permission from the BeingThere Centre.)

and other sensors need to be able to capture the 3-D geometry of everything inside a room and to project that representation as accurately as possible to the remote destination. "Not just the people, but the furniture, books, writing on the white board, waste basket, telephone, coffee mugs—everything," he says.

THE 3-D FUTURE

Three-dimensional technology has become one of the most active areas of telepresence research. Microsoft's Zhang is one of many researchers focusing on creating a 3-D virtual audio-visual environment that's both life-sized and photorealistic. "Such as that when you meet remote people in that environment, you feel as if they were sitting with you at the same table," he says. "This means you have the correct mutual gaze, accurate motion and stereoscopic parallax, and immersive spatial audio."

Zhang notes that his research is in a preliminary stage and is encountering the same road blocks that most advanced telepresence projects are bumping into: restricted Internet connection speeds and display resolution limits. "The current Internet bandwidth is not high enough, so we cannot send the full resolution of our entire 360° panoramic video," Zhang explains. "And even if we did send this in full resolution, a normal screen does not have enough resolution to display all of this."

Telepresence researchers also face a subtler challenge, one that most end users would have trouble defining, yet which gives many virtual conference participants the feeling that something just isn't "quite right." Telepresence experts call the problem gaze awareness: the ability to tell what someone is looking at by watching the direction of their eyes. Zhang says the challenge boils down to a matter of user perception. "People who are sitting at the same table see things differently: If I look at you, you see my face and your partner may see the side view, so your partner knows I am looking at you," he says. "If I turn my head and speak to your partner, then he sees my face, and you see the side view, and you know I'm not looking at you any more."

Conquering gaze awareness would go a long way toward improving telepresence realism and making sessions more like real world get-togethers. "In a virtual meeting, if you misinterpret a cue, you make the whole collaboration less effective," Zhang explains. "That's why, many times, people ask again and again for someone to clarify what's going on."

Nasser Peyghambarian, a professor of optical sciences at the University of Arizona in Tucson, believes that holography could provide the key to creating a truly immersive 3-D telepresence experience. "Holographic is closest to the way humans see their surroundings," he says. "It's also an approach that doesn't require any eyeglasses or other special eye wear, unlike when you go to see a 3-D movie or watch 3-D TV."

Peyghambarian notes that holographic presentation differs from today's movie and TV 3-D offerings in another important way. "If you go to *Avatar*, you see 3-D, but it has a very limited number of perspectives; something like two perspectives for one eye, and one for the other eye."

Holography, on the other hand, promises a vastly expanded number of perspectives, which would be handy for addressing challenges like gaze awareness. "Let's say it's a live object and there are 100 cameras taking 100 pictures from different angles," Peyghambarian says, "so there are 100 different perspectives that are coming in to provide detail." Using Peyghambarian's approach, the data generated by the camera array is

(continued on page 142)

Top Downloads in IEEE *Xplore*

This month marks the return of the "Reader's Choice" column. Each issue contains a list of articles published by the IEEE Signal Processing Society (SPS) that ranked among the top 100 most downloaded IEEE *Xplore* articles. This issue's column is based on download data through June 2011. The table below contains the citation information for each article and the rank obtained in IEEE *Xplore*. The highest rank obtained by an article in this time frame is indicated in bold. Your suggestions and comments are welcome and should be sent to Associate Editor Michael Gormish (gormish@ieee.org). [SP]

TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	ABSTRACT	RANK IN IEEE TOP 100 (2011) JUN	MAY	APR	MAR	FEB	JAN	*N* TIMES IN TOP 100 (SINCE JAN 2011)
AN INTRODUCTION TO COMPRESSIVE SAMPLING Candes, E.J.; Wakin, M.B. *IEEE Signal Processing Magazine,* vol. 25, no. 2, 2008, pp. 21–30	This article surveys the theory of compressive sampling, also known as compressed sensing or CS, a novel sensing/sampling paradigm that goes against the common wisdom in data acquisition.	18	43	40	31	54	**14**	6
WIRELESS SENSOR NETWORKS FOR COST-EFFICIENT RESIDENTIAL ENERGY MANAGEMENT IN THE SMART GRID Erol-Kantarci, M.; Mouftah, H.T. *IEEE Transactions on Smart Grid* vol. 2, no. 2, 2011, pp. 314–325	This paper evaluates the performance of an in-home energy management (iHEM) application comparing performance with an optimization-based residential energy management (OREM) scheme whose objective is to minimize the energy expenses of the consumers.	**34**						1
A TUTORIAL ON PARTICLE FILTERS FOR ONLINE NONLINEAR/NON-GAUSSIAN - BAYESIAN TRACKING Arulampalam, M.S.; Maskell, S.; Gordon, N.; Clapp, T. *IEEE Transactions on Signal Processing* vol. 50, no. 2, 2002, pp. 174–188	This paper reviews optimal and suboptimal Bayesian algorithms for nonlinear/non-Gaussian tracking problems, with a focus on particle filters. Variants of the particle filter are introduced within a framework of the sequential importance sampling (SIS) algorithm and compared with the standard EKF.	45	50	**32**	85		53	5
SUPER-RESOLUTION IMAGE RECONSTRUCTION: A TECHNICAL OVERVIEW Cheol Park, S.; Kyu Park, M.; Gi Kang, M. *IEEE Signal Processing Magazine* vol. 20, no. 3, 2003, pp. 21–36	This article introduces the concept of super resolutions (SR) algorithms and presents a technical review of various existing SR methodologies and models the low-resolution image acquisition process.	**50**						1
MULTIMEDIA CLOUD COMPUTING Wenwu Z.; Chong L.; Wang, J.; Shipeng L. *IEEE Signal Processing Magazine* vol. 28, no. 3, 2011, pp. 59–69	This article presents a multimedia-aware cloud to perform distributed multimedia processing and provide storage and quality of service (QoS) using the media-edge cloud (MEC) architecture.	58		93			**54**	3

Digital Object Identifier 10.1109/MSP.2011.942321
Date of publication: 1 November 2011

reader's CHOICE continued

TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	ABSTRACT	RANK IN IEEE TOP 100 (2011) JUN	MAY	APR	MAR	FEB	JAN	*N* TIMES IN TOP 100 (SINCE JAN 2011)
IMAGE SEGMENTATION USING FUZZY REGION COMPETITION AND SPATIAL/ FREQUENCY INFORMATION Choy, S.K.; Tang, M.L.; Tong, C.S. *IEEE Transactions on Image Processing* vol. 20, no. 6, 2011, pp. 1473–1484	This paper presents a multiphase fuzzy region competition model that takes into account spatial and frequency information for image segmentation. In the energy functional, each region is represented by a fuzzy membership function and a data fidelity term that measures the conformity of spatial and frequency data within each region to (generalized) Gaussian densities whose parameters are determined jointly with the segmentation process.	**65**						1
SMART TRANSMISSION GRID: VISION AND FRAMEWORK Li, F; Qiao, W.; Sun, H.; Wan, H.; Wang, J.; Xia, Y.; Xu, Z; Zhang, P. *IEEE Transactions on Smart Grid* vol. 1, no. 2, 2010, pp. 168–177	This paper presents a vision for the future of smart transmission grids. In this vision, each smart transmission grid is regarded as an integrated system that functionally consists of three interactive, smart components, i.e., smart control centers, smart transmission networks, and smart substations.	**74**						1
A HYBRID AC/DC MICROGRID AND ITS COORDINATION CONTROL Liu, X.; Wang, P.; Chiang Loh, P. *IEEE Transactions on Smart Grid* vol. 2, no. 2, 2011, pp. 278–286	This paper proposes a hybrid ac/dc micro grid to reduce the processes of multiple dc-ac-dc or ac-dc-ac conversions. Coordination control algorithms are proposed for smooth power transfer between ac and dc links and for stable system operation under various generation and load conditions.	**79**						1
IMAGE QUALITY ASSESSMENT: FROM ERROR VISIBILITY TO STRUCTURAL SIMILARITY Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. *IEEE Transactions on Image Processing* vol.13, no. 4, 2004, pp. 600–612	This paper introduces a framework for quality assessment based on the degradation of structural information. Within this framework a structure similarity index is developed and evaluated. MATLAB code available.	**84**						1
DIGITAL GRID: COMMUNICATIVE ELECTRICAL GRIDS OF THE FUTURE Abe, R.; Taoka, H.; McQuilkin, D. *IEEE Transactions on Smart Grid* vol. 2, no. 2, 2011, pp. 399–410	This paper introduces the "digital grid," dividing large synchronous grids into smaller segmented grids connected asynchronously via multileg IP addressed ac/dc/ac converters called digital grid routers. The digital grid can accept high penetrations of renewable power, prevent cascading outages, accommodate identifiable tagged electricity flows, record those transactions, and trade electricity as a commodity.	**94**						1
ADVANCES IN COGNITIVE RADIO NETWORKS: A SURVEY Wang, B.; Liu, K.J.R.; *IEEE Journal of Selected Topics in Signal Processing* vol. 5, no. 1, 2011, pp. 5–23	This paper surveys recent advances in research related to cognitive radios including: fundamentals of technology, network architecture, and applications. The existing works in spectrum sensing are reviewed along with important issues in dynamic spectrum allocation and sharing.	98		50	**27**	57	51	5
OFDM VERSUS FILTER BANK MULTICARRIER Farhang-Boroujeny, B. *IEEE Signal Processing Magazine* vol. 28, no. 3, 2011, pp. 92–112	The shortcomings of orthogonal frequency division multiplexing (OFDM) are examined in some applications and filter bank multicarrier (FBMC) is shown to be a potentially more effective solution.		**21**					1
SAMPLING RATE CONVERSION IN THE FREQUENCY DOMAIN [DSP TIPS & TRICKS] Guoan, B.; Mitra, S.K. *IEEE Signal Processing Magazine* vol. 28, no. 3, 2011, pp. 140–144	This article shows how to perform sampling rate conversion (SRC) for both integer and fractional-rate conversion by manipulating the discrete Fourier transform (DFT), implemented using the fast Fourier transform (FFT), of a time-domain signal.		**35**					1

TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	ABSTRACT	RANK IN IEEE TOP 100 (2011) JUN	MAY	APR	MAR	FEB	JAN	*N* TIMES IN TOP 100 (SINCE JAN 2011)
SMART GRID TECHNOLOGIES FOR AUTONOMOUS OPERATION AND CONTROL Meliopoulos, A.P.S.; Cokkinides, G.; Huang, R.; Farantatos, E.; Choi, S.; Lee, Y.; Yu, X. *IEEE Transactions on Smart Grid* vol. 2, no. 1, 2011, pp. 1–10	This paper presents a new smart grid infrastructure for active distribution systems that will allow continuous and accurate monitoring of distribution system operations and customer utilization of electric power.		70	81	**57**			3
THE SECURITY OF CLOUD COMPUTING SYSTEM ENABLED BY TRUSTED COMPUTING TECHNOLOGY Shen, Z; Tong, Q. 2010 2nd International Conference on Signal Processing Systems vol. 2, 2010, pp. V2-11–V2-15	Important security services, including authentication, confidentiality, and integrity, are provided by integrating the trusted computing platform into the cloud computing system.		82	78	40	**16**	28	5
SMART FAULT LOCATION FOR SMART GRIDS Kezunovic, M. *IEEE Transactions on Smart Grid* vol. 2, no. 1, 2011, pp. 11–22	This paper discusses issues associated with improving accuracy of fault location methods in smart grids using an abundance of data from intelligent electronic devices (IEDs).		**91**		94			2
FOCUS ON COMPRESSIVE SENSING [SPECIAL REPORTS] Edwards, J. *IEEE Signal Processing Magazine* vol. 28, no. 2, 2011, pp. 11–13	Discusses compressive sensing including single-pixel camera, MRI reconstruction, and some image representation.			53	**19**	100		3
MOBILE NETWORK SUPPORTED WIRELESS SENSOR NETWORK SERVICES Krco, S.; Tsiatsis, V.; Matusikova, K.; Johansson, M.; Cubic, I.; Glitho, R. IEEE International Conference on Mobile Adhoc and Sensor Systems, 2007, pp. 1–3	An architecture that utilizes the existing infrastructure to interconnect independent wireless sensor networks and to provide data aggregation and actuator control services is described.			83	**74**	86		3
LINEAR SUBSPACE LEARNING-BASED DIMENSIONALITY REDUCTION Jiang, X. *IEEE Signal Processing Magazine* vol. 28, no. 2, 2011, pp. 16–26	This article studies the linear subspace learning-based dimensionality reduction as a feature extraction module in the pattern recognition system.				**32**			1
DICTIONARY LEARNING Tošić, I.; Frossard, P. *IEEE Signal Processing Magazine* vol. 28, no. 2, 2011, pp. 27–38	This article describes methods for learning dictionaries that are appropriate for the representation of given classes of signals and multisensor data.				**43**			1
LEARNING LOW-DIMENSIONAL SIGNAL MODELS Carin, L.; Baraniuk, R.G.; Cevher, V.; Dunson, D.; Jordan, M.I.; Sapiro, G.; Wakin, M.B. *IEEE Signal Processing Magazine* vol. 28, no. 2, 2011, pp. 39–51	This article investigates the challenge of creating data models for dimensional reduction from the perspective of nonparametric Bayesian analysis.				**50**			1
SUBSPACE CLUSTERING Vidal, R. *IEEE Signal Processing Magazine* vol. 28, no. 2, 2011, pp. 52–68	This article presents a review of clustering high-dimensional data sets, including a number of existing subspace clustering algorithms together with an experimental evaluation on the motion segmentation and face clustering problems in computer vision.				**77**			1

reader's CHOICE continued

TITLE, AUTHOR, PUBLICATION YEAR IEEE SPS PUBLICATIONS	ABSTRACT	RANK IN IEEE TOP 100 (2011) JUN	MAY	APR	MAR	FEB	JAN	*N* TIMES IN TOP 100 (SINCE JAN 2011)
GEOMETRIC MANIFOLD LEARNING Jamshidi, A.A.; Kirby, M.J.; Broomhead, D.S. *IEEE Signal Processing Magazine* vol. 28, no. 2, 2011, pp. 69–76	This article describes algorithms and optimization criteria for analyzing massive and high dimensional data sets motivated by theorems from geometry and topology.				**87**			1
THE NEXT GENERATION OF POWER DISTRIBUTION SYSTEMS Heydt, G.T. *IEEE Transactions on Smart Grid* vol. 1, no. 3, 2010, pp. 225–235	This paper summarizes diverse concepts for the next generation of power distribution system in order to bring distribution engineering more closely aligned to smart grid philosophy.					79	**47**	2
FOR CLOUD COMPUTING, THE SKY IS THE LIMIT [SPECIAL REPORTS] Schneiderman, R. *IEEE Signal Processing Magazine* vol. 28, no. 1, 2011, pp. 15–144	A special report on cloud computing opportunities, issues, and surveys.					**88**	93	2
ACTIVE CONTOURS WITHOUT EDGES Chan, T.F.; Vese, L.A. *IEEE Transactions on Image Processing* vol. 10, no. 2, 2001, pp. 266–277	This paper proposes a model which can detect objects whose boundaries are not necessarily defined by the gradient, based on techniques of curve evolution, Mumford-Shah (1989) functional for segmentation and level sets.						63	1
FACE RECOGNITION BY EXPLORING INFORMATION JOINTLY IN SPACE, SCALE AND ORIENTATION Zhen Lei; Shengcai Liao; Pietikäinen, M.; Li, S.Z. *IEEE Transactions on Image Processing* vol. 20, no. 1, 2011, pp. 247–256	This paper proposes a face representation and recognition approach by exploring information jointly in image space, scale and orientation domains.						75	1
SIMPLIFIED RELAY SELECTION AND POWER ALLOCATION IN COOPERATIVE COGNITIVE RADIO SYSTEMS Liying Li; Xiangwei Zhou; Hongbing Xu; Li, G.Y.; Wang, D.; Soong, A. *IEEE Transactions on Wireless Communications* vol. 10, no. 1, 2011, pp. 33–36	This paper investigates joint relay selection and power allocation to maximize system throughput with limited interference to licensed (primary) users in cognitive radio (CR) systems.						82	1
REAL-TIME DEMAND RESPONSE MODEL Conejo, A.J.; Morales, J.M.; Baringo, L. *IEEE Transactions on Smart Grid* vol. 1, no. 3, 2010, pp. 236–242	This paper describes an optimization model to adjust the hourly load level of a given consumer in response to hourly electricity prices.						86	1
CLOCK SYNCHRONIZATION OF WIRELESS SENSOR NETWORKS Wu, Y.-C.; Chaudhari, Q.; Serpedin, E. *IEEE Signal Processing Magazine* vol. 28, no. 1, 2011, pp. 124–138	This article surveys the latest advances in the field of clock synchronization of wireless sensor networks (WSNs) from a signal processing viewpoint.						87	1
IMAGE SUPER-RESOLUTION VIA SPARSE REPRESENTATION Jianchao, Y.; Wright, J.; Huang, T.S.; Ma, Y. *IEEE Transactions on Image Processing* vol. 19, no. 11, 2010, pp. 2861–2873	This paper presents an approach to single-image superresolution, based upon sparse signal representation of low and high resolution patches.						90	1
DEEP LEARNING AND ITS APPLICATIONS TO SIGNAL AND INFORMATION PROCESSING [EXPLORATORY DSP] Yu, D; Deng, L. *IEEE Signal Processing Magazine* vol. 28, no. 1, 2011, pp. 145–154	This article introduces the emerging technologies enabled by deep learning and to review the research work conducted in this area that is of direct relevance to signal processing.						100	1

[from the GUEST EDITORS]

Touradj Ebrahimi, Lina Karam, Fernando Pereira, Khaled El-Maleh, and Ian Burnett

The Quality of Multimedia: Challenges and Trends

One of the most visible consequences of progress in multimedia has been an increase in quality of the products and services offered. Nowadays, there is no longer only a question of which features are to be included in multimedia applications but also how well they contribute to the quality of user experience.

But what is meant by quality in the context of multimedia signal processing and its applications, especially how should it be measured in a meaningful, reliable, and reproducible way?

Quality is one of those fundamental notions in science reaching back several centuries. One can find traces of it in the works of Aristotle, where he categorized every object of human apprehension into ten categories, one of which was quality. Quality has long been among the important performance measures in communication and information technologies as witnessed by the huge work surrounding quality of service. More recently, as user-centric multimedia has become increasingly important, interest in quality (and more specifically, quality of experience) has also increased, taking into account not only content fidelity metrics and system performance parameters but also encompassing notions such as user perception, user acceptance, context, and expectation.

This special issue of *IEEE Signal Processing Magazine* aims to provide an overview of the state of the art in quality issues surrounding multimedia signal processing and to highlight challenges and trends in this field. The following eight articles cover a large, but by no means complete, spectrum of topics addressing quality in various modalities ranging from speech to image and video processing and coding as well as interactions between them.

In "Speech Quality Estimation," Möller et al. present a tutorial overview of speech quality estimation models with a focus on quality of user experience. After surveying existing state-of-the-art signal-based, parametric, information-based, and hybrid models, they provide practical guidance on how to select the best quality assessment model for a given voice application. The article concludes by highlighting some of the limitations of existing models and ongoing research efforts to develop new and better alternatives.

In the article by Wang and Bovik, "Reduced- and No-Reference Image Quality Assessment," the authors highlight the need for reduced- and no-reference quality assessment metrics in today's image and video applications, as opposed to full-reference approaches. The authors then give a tutorial overview of the problem by clarifying the issues to be solved and how they might be pragmatically addressed and provide insights in trends and challenges that remain.

"Video Is a Cube," by Keimel et al., inspired by notions from chemometrics, introduces an innovative design of video quality metrics via data analysis. In particular, it employs multidimensional data analysis for a better exploitation of higher-dimensional data in video quality metrics, hence allowing to more properly take into account temporal properties of video content.

In "Visual Attention in Quality Assessment," Engelke et al. provide a review of recent advances in visual attention in the context of quality assessment and discuss remaining challenges in addition to potential solutions and future directions in this field.

"Audiovisual Quality Components," by Pinson et al., examines the relative importance and influence of audio and video quality in an audiovisual sequence. The first part of the article examines an audiovisual experiment that looks to devise a generally applicable regression model predicting audiovisual quality. The remainder of the article compares the results with a dozen previous experiments, justifying a generalized cross-term model of audiovisual quality.

"IP-Based Mobile and Fixed Network Audiovisual Media Services," by Raake et al., provides a survey of current approaches for monitoring the quality perceived by users in IP-based audiovisual media services over both fixed and mobile networks. A large portion of the article is dedicated to the review of quality estimation models that exploit available information in media packets to help in assessing quality of service. The last part of the article presents a summary of some of the emerging trends and challenges related to media service monitoring.

In their article "Assessing Visual Quality of 3-D Polygonal Models," Bulbul et al. review recent advances in evaluation and measurement of the perceived visual quality of three-dimensional (3-D) polygonal models. This article analyzes the processing steps of objective quality assessment metrics and subjective user evaluation methods and presents a taxonomy of existing solutions. In this process, the article discusses existing metrics, including perceptually based ones, computed either on 3-D data or on two-dimensional projections, and

(continued on page 148)

Digital Object Identifier 10.1109/MSP.2011.942546
Date of publication: 1 November 2011

Sebastian Möller, Wai-Yip Chan, Nicolas Côté,
Tiago H. Falk, Alexander Raake, and Marcel Wältermann

Speech Quality Estimation

[Models and trends]

This article presents a tutorial overview of models for estimating the quality experienced by users of speech transmission and communication services. Such models can be classified as either parametric or signal based. Signal-based models use input speech signals measured at the electrical or acoustic interfaces of the transmission channel. Parametric models, on the other hand, depend on signal and system parameters estimated during network planning or at run time. This tutorial describes the underlying principles as well as advantages and limitations of existing models. It also presents new developments, thus serving as a guide to an appropriate usage of the multitude of current and emerging speech quality models.

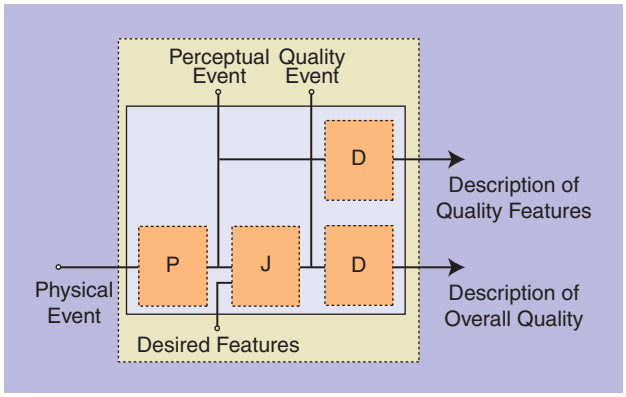
INTRODUCTION

Since the large-scale introduction of telephony networks, efforts have been made to guarantee high-quality and reliable services to human users. Transmission performance was initially measured by informally exchanging phonetically rich phrases between two network terminals, thus quantifying the intelligibility associated with the channel. Later, such informal procedures were replaced by standardized listening-only and conversational tests that provided more stable conditions—and thus smaller confidence intervals—when asking test participants to rate the (perceived) loudness or intelligibility, listening effort, or overall quality of the heard speech samples or conversations [28].

For a participant in a quality judgment experiment, for example, speech quality is regarded as a multidimensional construct, as it is the result of three processes: perception (P), judgment (J), and description (D), as depicted in Figure 1. The perception processes are triggered by a so-called "physical event" (i.e., a sound wave reaching the human ears), which gives rise to a "perceptual event." We use the term "event" to denote an instance of occurrence of a phenomenon in time and space; see [4]. This "perceptual event" can also be described in a multidimensional way, wherein

Digital Object Identifier 10.1109/MSP.2011.942469
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE



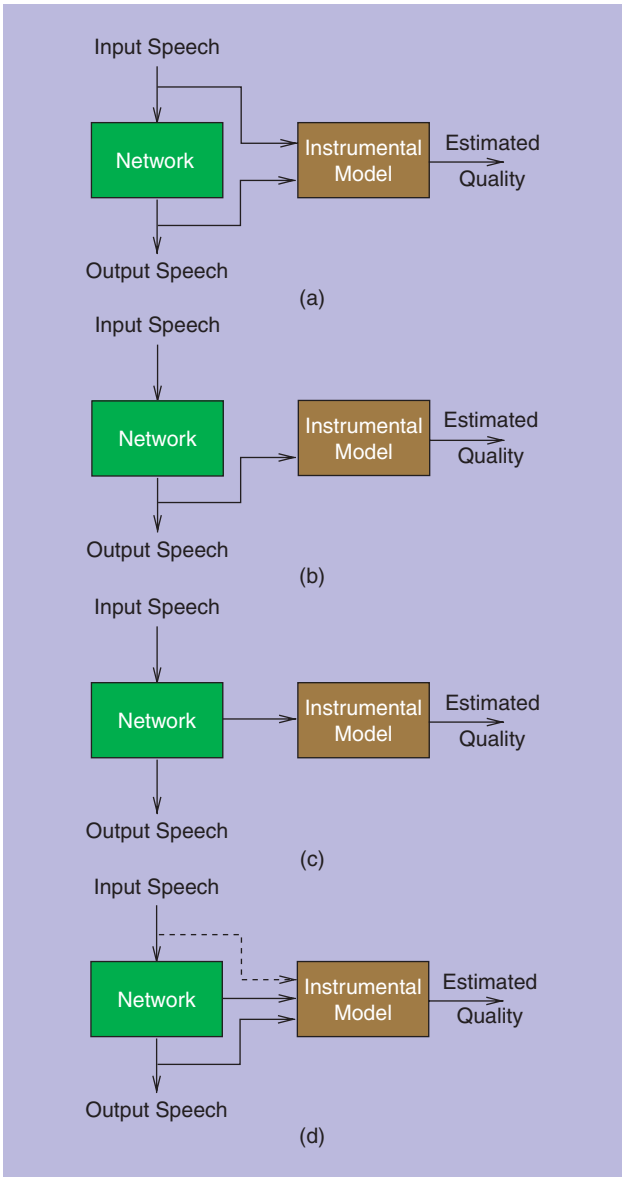
[FIG1] Schematic representation of a participant in a quality judgment experiment; see [55].

features such as loudness, coloration, or noisiness are quantified [55], [60], [61].

The features of the perceptual event are further compared to the desired features of some internal reference [42], [55]. This reference can be formed via repeated telephone usage experiences, but it also reflects numerous context- and situation-dependent factors such as the user's expectations (e.g., free versus paid call), motivation (e.g., urgent call), and experience (e.g., avid mobile phone user); the test setup (e.g., listening-only, conversational) and audio bandwidth (e.g., narrowband versus wideband); as well as environmental factors (e.g., noisy versus quiet environments), to name a few [51]. The result of this comparison is a "quality event" that may be quantitatively described as a judgment of the "overall quality." Unfortunately, both the "perceptual event" and the "quality event" are internal to the perceiving human (denoted by the dotted line around the processes in Figure 1). To quantify salient attributes of these internal events, one has to rely on human test participants expressing their "subjective" judgments in terms of opinion scores. Most commonly, the mean opinion score (MOS) is used where the individual participants' scores are averaged to level out individual factors. As such, subjective methods are time-consuming, laborious, and expensive, thus prompting the development of instrumental or so-called "objective" speech quality estimation methods.

Instrumental models are used to estimate the average user judgment of the quality of a service. Commonly, though not necessarily, individual experiences and requirements are not taken into account by these models. Most models provide an estimate of the "overall quality" judged in a quiet listening-only or conversational context according to standardized test conditions [34], or with the consideration of background noise and its suppression [37]. Other models estimate multiple quality features such as coloration, noisiness, and continuity [61], [7]. The models (Figure 2) base their estimations

1) on signals that can either be measured at the electrical or the acoustic interfaces of the transmission channel of interest (signal based)



[FIG2] Types of instrumenal models. (a) Full-reference signal-based model; (b) reference-free signal-based model; (c) parametric model; (d) protocol-information-based and hybrid model.

2) on parameters that are estimated during the network planning phase (parametric)

3) on parameters collected at run time from network processes and control protocols

4) on a combination of 1) and 2) (hybrid models).

In the last scenario, the speech signal at the network output can be used alone or together with the network input speech signal, as depicted by the dashed line in Figure 2(d). Existing models can provide quality estimates for the classical narrowband (NB) (300–3,400 Hz), wide-band (WB) 50–7,000 Hz), or superwide-band (SWB) 50–14,000 Hz) signals. Table 1 gives an overview of models currently standardized or being discussed by relevant standardization bodies.

SIGNAL-BASED MODELS

Signal-based models employ speech signals transmitted or otherwise modified by speech processing systems to estimate quality. Most models provide quality estimations according to the Absolute Category Rating (ACR) listening quality scale defined in [34], but recently other models have been designed to predict individual quality features [52], [58], [62]. Two types of signal-based models exist: full-reference (also known as "intrusive" or "double-ended") models, which depend on a reference (system input) speech signal and a corresponding degraded (system output) speech signal; and reference-free ("nonintrusive" or "single-ended") models, which depend only on the latter degraded signal.

TWO TYPES OF SIGNAL-BASED MODELS EXIST: FULL-REFERENCE AND REFERENCE-FREE MODELS.

The idea of a full-reference model for predicting listening-only quality is simple—assuming that the aim is to transmit a speech signal over a channel without any perceptual degradation, then the perceptually weighted distance between the channel input and output signals should be indicative of the speech transmission quality. It is important that the distance be calculated on a perceptual level, as modern speech transmission channels do not aim at reproducing the exact signal, but only generate a similarly sounding signal at the output.

This underlying principle also points at some principal weaknesses of such models; because a comparison is made with respect to the input signal, this signal also has to reflect all the desired features of Figure 1 to correctly reflect the quality judgment process. Most models, however, predict the judgment made in an ACR test, and not in a paired-comparison paradigm. In an ACR test, the listener has no direct access to the reference input signal; the desired features are induced from the listener's experience by the test context, e.g., by the fact that the test contains only NB, WB, or SWB stimuli. The test context circumscribing the listener judgment is usually accounted for by selecting a model usage mode (NB, WB, or SWB) or separate model varieties for different test contexts.

Existing reference-based models comprise three components: 1) a preprocessing step including a level- and time-alignment of the two speech signals; 2) a perceptual transformation of the speech signals simulating parts of the peripheral human auditory system; and 3) an assessment unit that compares the two perceptually transformed signals. The most widespread full-reference model, the Perceptual Evaluation of Speech Quality (PESQ) [39] provides quality estimates for NB speech signals. It is based on its predecessor, the Perceptual Speech Quality Measure [(PSQM), formerly standardized in [38], but shows an improved performance for packet-switched networks by employing a better time-alignment algorithm and a different perceptual model]. To support burgeoning WB speech services, a WB extension of PESQ, called WB-PESQ [40], was standardized in 2005. However, this model has a limited scope, as it does not cover electroacoustic transducers, voice quality enhancement (VQE), and time-warping algorithms [39]. Therefore, Study Group 12 of the International Telecommunication Union (ITU) developed a new model called Perceptual Objective Listening Quality Assessment (POLQA) [24], which provides quality estimations in both NB and SWB contexts, and which is intended to cover the majority of existing telephone network scenarios.

The POLQA model provides quality estimation for fixed, mobile, and IP-based telephony services, including speech processing systems such as G.711.1, G.718, Skype SILK, Adaptive Multirate AMR-WB+, Advanced Audio Coding AAC LD, Enhanced Variable Rate Codecs (EVRC), and Continuous Variable Slope Delta Modulation (CVSD) codecs, as well as VQE algorithms (e.g., noise reduction, bandwidth extension, and automatic gain control). In its SWB mode, POLQA has a

[TABLE 1] TAXONOMY OF STANDARD OBJECTIVE SPEECH QUALITY PREDICTION MODELS.

QUALITY ASPECT	TYPE OF INPUT	INPUT	AUDIO BANDWIDTH	EXAMPLE
L-OQ	SIGNAL	M: 1 e	NB	P.563 [32], ANIQUE+[46]
		M: 2 e	NB	PSQM (P.861) [38], PESQ (P.862) [39]
			WB	WB-PESQ (P.862.2) [40]
		M: 2 e/a	NB/SWB	P.OLQA (P.863) [24]
	PARAM.	P	NB, WB	P.564 [33]
L-N	SIGNAL	M: 2e	NB	ETSI EG 202 396-3 [10]
L-M	SIGNAL	M: 2e/A	NB/SWB	P.AMD [25]
C-OQ	SIGNAL	M: 1e	NB	P.562 (CALL CLARITY INDEX [31], NONINTRUSIVE E-MODEL)
		M: 2e	NB	PESQM [2]
	PARAM.	E	NB	E-MODEL (G.107) [29]
			WB	WB E-MODEL (G.107) [29]
C-M	PARAM.	E	WB	DIMENSION-BASED WB E-MODEL [27]

Quality aspects: L = listening-only overall quality; L-N = listening-only for speech quality, noise quality, and overall quality; L-M = listening-only with several quality dimensions; C-OQ = conversational overall quality; C-M = conversational with several quality dimensions. Input: M = measurement; P = protocol information; E = offline measurement or estimation; 1 = one signal; 2 = two signals; a: acoustic signal, i.e., talker's speech signal measured by microphone(s); e: electric signal, i.e., talker's speech signal measured anywhere along the network transmission path. Audio bandwidth: NB = 300–3,400 Hz; WB = 50–7,000 Hz; SWB = 50–4,000 Hz. Exemplary models will be discussed in the subsequent text.

wider scope than WB-PESQ: it covers a wider bandwidth, from NB to SWB, and specific degradations such as frequency distortions introduced by user's terminal and nonoptimal listening levels. The assessment unit in POLQA uses an "idealized" signal, instead of the standard input reference signal for comparison; computes six different quality values; and combines them into an overall speech quality estimate. The estimate is computed in the so-called "cognitive" model that simulates high-level cognitive processes. Details on the standardized model can be found in [24].

REFERENCE-FREE MODELS HAVE GAINED MUCH ATTENTION RECENTLY AS REFERENCE SPEECH SIGNALS ARE NOT READILY AVAILABLE WITH IN-SERVICE NETWORKS.

Signal-based models for assessing the overall quality of a transmission channel provide an estimate of the quality level that can be reached with that particular channel; however, they do not provide insight into *why* a particular channel is good or bad. Thus, further information is necessary to diagnose the sources of poor quality. Such insight can be gained from models that predict multiple quality features. For a variety of modern transmission channels, Wältermann et al. [61] have uncovered three underlying orthogonal perceptual dimensions, i.e., discontinuity, coloration, and noisiness. Additionally, as recently documented in [7], the inclusion of a loudness dimension can also be useful in cases where nonoptimal (i.e., too high or too low) listening levels are present. Other dimensions pertaining to signal and/or background perceptual quality have been proposed in [60]. These dimensions were first determined via psychoacoustic experiments. Subsequently, multidimensional instrumental reference-based quality models were developed to estimate such percepts, such as the four-dimensional model recently documented by Côté [8], among others (e.g., [52], [58], and [59]). Moreover, on the basis of multiple computed quality features, it is possible to estimate the overall quality as a linear combination of the constituent features, as proposed in [62].

Estimation of multiple dimensions is also advantageous when characterizing the quality of noise-suppressed speech. Noise suppression algorithms can introduce unwanted artifacts to the speech signal, such as musical noise. In such situations, listeners can become confused as to which components of a noisy speech signal should form the basis of their ratings of overall quality. To reduce the error variance (or listener uncertainty) in the subjects' ratings of overall quality, the subjective test procedure recommended in ITU-T Rec. P.835 [37] instructs the listener to successively attend to and rate three different components of the noise suppressed speech signal: the speech signal alone (SMOS), background noise alone (NMOS), and the overall quality effect (GMOS). A full-reference model that estimates these three indices has been developed by the European Telecommunications Standardization Institute (ETSI) and is recommended in [10]. Besides an estimation of the speech quality that is based on a clean speech reference signal, the noise impact is estimated by means of the so-called "relative approach" [17], which emphasizes dynamic characteristics of the noise signal.

Reference-free models have gained much attention recently as reference speech signals are not readily available with in-service networks. Like the participants in ACR-type listening tests, reference-free models assess speech quality without the need to "listen" to a high-quality "clean" version of the target signal. Humans, through their experience, have acquired knowledge of normal and abnormal phenomena in speech sounds. Subjects in ACR listening tests rely only on this prior knowledge (desired features in Figure 1) to judge speech quality. In a similar vein, reference-free models utilize prior knowledge of normal and/or abnormal behavior of speech signal features to estimate speech quality. To represent prior knowledge, existing reference-free models employ models of speech production [19], speech perception [46], speech signal feature likelihood [14], [43], or a combination thereof [48].

In 2004, ITU-T held a competition to standardize a nonintrusive signal-based measure. Two algorithms stood out during this competition; one became the ITU standard P.563 [32] and the other, ANIQUE+, became an American National Standard Institute (ANSI) standard [1], [46]. While these models have been shown to be reliable for many telecommunications scenarios, recent research has suggested that their quality prediction performance is compromised for scenarios involving VQE algorithms (e.g., noise suppression [9] and dereverberation [16]) and wireless-VoIP tandem connections [12]. For both NB and WB reverberant and dereverberated speech, a no-reference quality model termed speech-to-reverberation modulation energy ratio (SRMR) has been recently proposed [16] and is available as open-source software for academic and research purposes. Directives on how to download the SRMR toolbox for MATLAB (Mathworks) can be obtained by contacting Tiago H. Falk. Table 2 summarizes application conditions in which the abovementioned standardized signal-based quality models have been recommended to be used and to be avoided.

The signal-based models presented so far aim at predicting speech quality or its features in a listening-only context. However, a primary goal of telecommunication speech services is to enable users to interact through conversations. Ease of conversation or interaction is however not measured in listening-only quality assessments, where the listeners do not interact with the speaker. In telephony conversations, interactivity is affected by the mouth-to-ear transmission delay. Large delays make it hard to interrupt the speaker, to promptly exchange turn for speaking, and for two simultaneous speakers to quickly return to a single-talker configuration. Also, the annoyance level of echoes [18] increases with delay. Echoes can be caused by reflections of the talker's speech across four-wire-two-wire circuit interfaces or across the acoustic interface (loudspeaker-microphone coupling).

The ITU-T P.561 standard specifies requirements for in-service nonintrusive measurement devices (INMDs) to measure two-way speech transmission path parameters such as speech and noise levels, echo loss, and path delay [30]. A proprietary method called call clarity index (CCI), described in ITU-T P.562 Annex A, maps these measured parameters to an estimated conversational MOS value [31]. Within this paradigm, researchers have devised signal measurement algorithms to calculate the "planning" parameters in the parametric E-model (described in the section "Parametric Models") for the purpose of estimating listening or conversational quality.

APART FROM PARAMETRIC MODELS LIKE CCI AND THE E-MODEL, SIGNAL-BASED MODELS ALSO HAVE BEEN EXTENDED TO PROVIDE ESTIMATIONS OF CONVERSATIONAL QUALITY.

Apart from parametric models like CCI and the E-model, signal-based models also have been extended to provide estimations of conversational quality. In a first step, Appel and Beerends [2] developed a model for talking-only speech quality, called Perceptual Echo and Sidetone Quality Measure (PESQM), which is based on PESQ. It takes into account the impact of degradations such as talker echo, coding artifacts, and additive noise on the talker's perception of his/her own voice. Guéguin et al. [20] proposed a signal-based model of conversational speech quality that combines both PESQ and PESQM with an estimation of the conversational impact of pure delay, the latter being derived from the E-model. None of these approaches has been standardized yet, but the definition and validation of a signal-based model for conversational speech quality is a work item in ITU-T Study Group 12.

PARAMETRIC MODELS

Signal-based models require speech signals as input to the quality estimation method. Thus, at least a prototype implementation or simulation of the transmission channel has to be set up. During the network design process, such signals are commonly not available; instead, the network is characterized by the technical specifications of its constituent elements. Such specifications include, amongst others, the frequency-weighted insertion loss (so-called "loudness rating") and the delay associated with a particular transmission path, the power of signal-correlated or uncorrelated noise inserted by the equipment, the probability that packets get lost or discarded in Internet-Protocol-(IP)-based transmission, as well as the type of speech codec and error concealment techniques used. Most of these specifications can be quantified in terms of planning parameters that enable parametric estimation of speech quality prior to the connection becoming live.

The E-model [44] can be treated as an archetypal parametric model used to estimate the quality associated with a speech transmission channel in a conversational context. Thus, in contrast to models that estimate listening quality, the E-model takes "two-way interaction effects" such as delay and echoes into account. The model features impairment factors that parametrically capture the different types of impairments in a telephone connection, covering the complete transmission chain from the mouth of the speaker to the ear of the listener.

The E-model impairments are grouped into four classes: 1) impairments affecting the basic signal-to-noise ratio of the transmission channel, such as ambient noise or circuit noise; 2) impairments occurring simultaneously with the speech signal, such as a nonoptimal sidetone level, or quantization distortions resulting from pulse-code modulation (PCM); 3) impairments occurring delayed with respect to the speech signal, such as talker and listener echoes, or the conversational impact of pure delay; and 4) impairments resulting from nonlinear and/or time-varying equipment, such as coding distortions or the effects of packet loss.

[TABLE 2] APPLICATION CONDITIONS IN WHICH EXISTING SIGNAL-BASED MODELS ARE RECOMMENDED TO BE USED AND TO BE AVOIDED.

MODEL	RECOMMENDED FOR	LIMITATIONS
PESQ	INPUT LEVELS, TRANSMISSION CHANNEL ERRORS, PACKET LOSS WITH OR WITHOUT CONCEALMENT, BIT-RATES, TRANSCODINGS, NOISE AT SENDING SIDE, TIME-VARYING DELAY, WAVEFORM CODECS, CELP CODECS, OTHER CODECS (CF. [39])	LISTENING LEVELS, LOUDNESS, HYBRID CODECS SUCH AS AMR AND EVRC, TIME-WARPING, NOISE REDUCTION, ECHO CANCELLATION
WB-PESQ	AS ABOVE, BUT WITH WB TRANSMISSION	AS ABOVE WITH WB TRANSMISSION, HYBRID CODECS SUCH AS AMR-WB, G.729.1 AND EVRC-WB (CF. [8])
POLQA	SAME AS PESQ ABOVE, BUT WITH SWB TRANSMISSION, VQE ALGORITHMS, SHORT TIME-WARPING ALGORITHMS, ELECTRO-ACOUSTIC TRANSDUCERS, HYBRID SPEECH CODECS	STRONG TIME-WARPING DISTORTIONS, EVRC CODECS
ETSI EG 202 396-3	NOISE REDUCTION ALGORITHMS, IN NB OR WB TRANSMISSIONS	
P.563	SAME AS PESQ ABOVE AND FOR SHORT- AND LONG-TERM TIME WARPING OF THE SPEECH SIGNAL (CF. [32])	TALKER ECHO, SIDETONES, LOW BITRATE (< 4KBPS) LPC VOCODER TECHNOLOGIES, SINGING VOICE. ALSO, APPLICATION SCENARIOS INVOLVING VQE ARTIFACTS, BITRATE MISMATCH BETWEEN ENCODER AND DECODER, AND AMPLITUDE CLIPPING WERE NOT FULLY VALIDATED DURING THE TIME OF THE STANDARDIZATION
ANIQUE+	SAME AS P.563 ABOVE	SAME AS P.563 ABOVE

In each class, the degradation is quantified in terms of a so-called "impairment factor" that is assumed to be additive on a perceptual scale. Thus, the overall transmission rating of the connection can be expressed by

$$R = Ro - Is - Id - Ie, eff + A, \quad (1)$$

where Ro is the basic signal-to-noise ratio of the channel, Is is the impairment factor related to the simultaneous degradations, Id is the impairment factor related to delayed degradations, and Ie, eff is the impairment factor related to nonlinear and time-varying degradations. A is called the advantage factor and reflects the quality expectation of the user. Depending on particular circumstances, such as mobile connections or connections to hard-to-reach areas, the user's quality expectation may differ from the norm; roughly speaking, A serves to adjust the desired features of Figure 1. The final transmission rating R (range: 0 = worst… 100 = best) can easily be transformed to a conversational MOS, which is the average rating on an overall quality scale collected in a conversation test carried out according to [34], following an S-shaped curve defined in [29].

Note that the transmission rating scale R can also be deployed for providing information on a purely perceptual level. In other words, the transmission system-based impairment factors described above can be replaced by multiple perceptual-dimension-related impairment factors, such as discontinuity, noisiness, and coloration [27].

Also, the E-model was originally developed for NB speech transmission. With an increase in wideband VoIP usage, the E-model framework has recently been updated to also provide valid predictions for WB transmission [54]. A complete WB-version of the E-model is expected at the end of the current ITU-T study period (2009–2012). In this case, the transmission rating scale is extended to the range 0…129 to reflect the potential quality advantage of WB transmission. Further extensions are currently being discussed for SWB transmission, where maximum values of $R = 179$ have been observed in auditory tests [64]. Due to its general nature and dispensation of signal input, the E-model is especially attractive for network planning. For this such purpose, the E-model is recommended by the ITU-T [29]. A list of recommended and not recommended application conditions of the NB and the WB version of the E-model is given in Table 3.

PROTOCOL-INFORMATION-BASED MODELS

The E-model has also been used for monitoring quality of VoIP in many studies [55], but it often does not provide accurate measurements for individual calls. As a consequence, alternative models based on protocol information as input have been developed. Instead of using the voice payload of the transmitted packets, the models exploit protocol header information such as the time stamps and sequence numbers from RTP [22] headers for delay and packet-loss-related information, and information on the end-point behavior such as dropped packets statistics or PLC information [23]. The main goal of such models is to enable passive network and/or end-point monitoring with a lightweight parametric approach, at the same time avoiding privacy concerns when accessing user-related payload information. The models can be employed at different locations in the service chain; by locating the model and measurement points in the client, the network, or both, solutions adaptable to different architectures are enabled (see the section "Practical Guidance" for more details). Examples of models of this type are described in [5] and [6]. Instead of standardizing an individual method, ITU-T recommends a procedure for validating NB- or WB-listening quality monitoring models by taking PESQ predictions as the reference; this procedure is described in [33]. An update of this procedure is expected with the new POLQA standard [24] for full reference quality prediction. One of the main differences to parametric planning models such as the E-model is that quality is followed and pooled over time, enabling more accurate predictions of per-call quality for the case of nonuniform packet loss. A comparison of parameter-based models employing different ways of temporal integration or pooling can be found, for example, in [53]. Table 4 summarizes the application areas of models that correspond to [33].

HYBRID APPROACHES

Modern communication scenarios can involve tandeming and internetworking of heterogeneous links, thus leading to impairment combinations that compromise the performance of existing quality measurement algorithms. For instance, in

[TABLE 3] APPLICATION CONDITIONS IN WHICH PARAMETRIC MODELS ARE RECOMMENDED TO BE USED AND TO BE AVOIDED.

MODEL	RECOMMENDED FOR	LIMITATIONS
E-MODEL	NB HANDSET TELEPHONY, INCLUDING THE EFFECTS OF OVERALL LOUDNESS, FREQUENCY DISTORTION, QUANTIZING DISTORTION, CODING, BACKGROUND NOISE, CIRCUIT NOISE, NONOPTIMUM SIDETONE LEVEL, TALKER AND LISTENER ECHO, PURE DELAY, RANDOM AND BURSTY PACKET LOSS	DIFFERENT LISTENING LEVELS, NONHANDSET TERMINALS INCLUDING NOISE REDUCTION AND ECHO CANCELLATION [50]
WB E-MODEL	WB TELEPHONY WITH HANDSET AND HEADPHONE LISTENING, INCLUDING CODING DISTORTIONS AND RANDOM PACKET LOSS [29]; FIRST EXTENSIONS PROPOSED FOR OVERALL LOUDNESS, FREQUENCY DISTORTION, AND SEND SIDE NOISE	NONOPTIMUM SIDETONE LEVEL, TALKER AND LISTENER ECHO, PURE DELAY, OTHER TERMINALS INCLUDING NOISE REDUCTION AND ECHO CANCELLATION, BANDWIDTH EXTENSION ALGORITHMS

[TABLE 4] APPLICATION CONDITIONS IN PROTOCOL-INFORMATION-BASED MODELS ARE RECOMMENDED TO BE USED AND TO BE AVOIDED.

MODEL	RECOMMENDED FOR	LIMITATIONS
P.564	APPLICATION RANGE DETERMINED BY INPUT INFORMATION AVAILABLE AT MEASUREMENT POINT AND APPLICATION RANGE OF PESQ	PREDICTIONS FOR EFFECTS NOT ADDRESSED BY PESQ, SUCH AS ECHO OR DELAY, CANNOT BE VALIDATED USING [33]; ECHO AND OTHER INFORMATION AS AVAILABLE E.G., FROM THE CLIENT [23] MUST BE MEANINGFUL

wireless VoIP tandem communications, standard signal-based models such as PESQ [39] and P.563 [32] were shown to be sensitive to varying packet loss rates and PLC strategies [12], [13]. Parametric models, in turn, were shown to be sensitive to acoustic background noise combined with PLC artifacts, as well as noise suppression artifacts combined with speech codec distortions [13].

Hybrid approaches, which can make use of both the signal decoded from the payload and IP connection parameters extracted from protocol header information, have been proposed to overcome the limitations of pure signal and pure parametric/protocol-based approaches. The model developed in [13], for example, made use of IP connection parameters such as codec and PLC type, packet size, and packet loss pattern to determine a "base quality" representative of the transmission link under test. Distortions that were not captured by the connection parameters, such as the acoustic noise type and level, temporal clippings, and PLC and noise suppression artifacts, were computed from perceptual features extracted from the decoded speech signal and used to adjust the base quality accordingly.

Alternately, hybrid approaches can also be taken when input parameters (e.g., the codec algorithm used to generate the speech payload or other codec-related impairment factors) of a parametric model are unobserved or otherwise unavailable. For example, if the E-model is to be used in conjunction with a new type of codec, a corresponding effective equipment impairment factor *Ie, eff* has to be derived. For this purpose, the ITU-T recommends either to carry out listening-only tests, or to rely on signal-based full-reference models. A full-reference model such as PESQ is able to estimate adequate *Ie, eff* values, provided that the estimations are normalized for biases resulting from the test stimuli; the corresponding normalization procedures are defined in [35] for NB and in [36] for WB speech transmission.

In general, hybrid approaches entail fusing diverse information that is more readily or economically accessible, reliable, or timely in specific speech quality estimation applications. For instance, in speech transmission over tandem links, measuring the degradations at each link endpoint and distributing the measurements across network nodes along the transmission path provides more accurate estimates of speech quality than relying only on information at transmission endpoints. Such distributed measurement also provides diagnostics that would enable isolating sources of degradation to specific network elements. With evolving wireline and wireless networks, opportunities abound for embedding into network elements functionalities for distributed quality monitoring, diagnosis, remedy, and assurance.

PRACTICAL GUIDANCE

Given the multitude of available models, it is not always apparent to the practitioner which model to apply for a given purpose. To further complicate the situation, users are often faced with multiple models that are deemed applicable to a specific application. In Table 5, we compile a list of currently recommended or in-use approaches for speech quality prediction, including emerging standards which are still under discussion. It is hoped that the information provided in the table and documented below will assist users and/or researchers in selecting an appropriate model for their specific application.

For network planning purposes, the only currently recommended model is the E-model. Its NB version covers all standard network elements as well as standard handset terminals; first extensions for nonhandset terminals including signal-processing equipment such as noise reduction and echo cancellation have been presented [45], [50], but they are not yet conclusive. For WB, the underlying scale has been extended and *Ie,eff* values are provided in [29]; other types of degradations have been dealt with in [54], but they are not yet recommended by the ITU-T. For SWB, only a transmission rating scale extension has been proposed [64]. First proposals have also been made to predict individual quality features with a perception-based E-model, but they are not yet conclusive [27]. Users should exercise care when considering the nonconcluded elements in their applications and/or research.

For network optimization and maintenance, the upcoming recommended model is the reference-based P.OLQA, published as ITU-T Rec. P.863 in early 2011 [24]. It comes in NB and SWB modes, the latter also covering WB scenarios and acoustic recording conditions. P.863 will, perhaps gradually, replace NB [39] and WB P.862 PESQ [40], which are still widely used in field measurement equipment as well as in the laboratory. P.863 has shown significantly better performance than P.862 on a large variety of databases and a wider range of usage scenarios [26]. Current work in ITU-T SG12 is focusing on the characterization of the new model, as well as on the development of a multidimensional model Perceptual Approaches for Multidimensional (P.AMD) analysis [25], which is able to estimate four to seven quality features, describing discontinuity,

[TABLE 5] APPLICATION SCENARIOS OF SPEECH QUALITY PREDICTION MODELS.

APPLICATION	TYPE OF INPUT	SOURCE OF INPUT	TARGET QUALITY PREDICTION	TARGET AUDIO BANDWIDTH	CONSIDERED CHANNEL ELEMENTS AND SCOPE	RECOMMENDED MODEL
PLANNING	PARAMETERS	ESTIMATION	MOS-CQEN	NB	ALL MOUTH-TO-EAR	G.107 [29]
	PARAMETERS	ESTIMATION	MOS-CQEM	WB	ALL MOUTH-TO-EAR	G.107 [29] + EXTENSIONS [54] (SEE NOTE 1 BELOW)
	PARAMETERS	ESTIMATION	MOS-CQEM	SWB	ALL MOUTH-TO-EAR	G.107 [29]+ EXTENSIONS [64] (SEE NOTE 1 BELOW)
	PARAMETERS	ESTIMATION	QUALITY FEATURES	SWB	ALL MOUTH-TO-EAR	G.107 [29] + EXTENSIONS [27] (SEE NOTE 1 BELOW)
OPTIMIZATION	2e SIGNALS	SIMULATION	MOS-LQON	NB	CODECS, TRANSMISSION ERRORS, NOISES AT THE SENDING SIDE	P.862 [39] (SEE NOTE 2 BELOW)
	2e OR 2a SIGNALS	SIMULATION	MOS-LQON	NB	CODECS, TRANSMISSION ERRORS, TERMINALS, TIME-WARPING, ETC.	P.863 [24]
	2e SIGNALS	SIMULATION	MOS-LQOW	WB	SAME AS P.862	P.862.2 [40] (SEE NOTE 2 BELOW)
	2e OR 2a SIGNALS	SIMULATION	MOS-LQOM	SWB	CODECS, TRANSMISSION ERRORS, TERMINALS, TIME-WARPING, NOISES AT THE SENDING SIDE, LISTENING LEVELS	P.863 [24]
	3e SIGNALS	SIMULATION	SMOS, NMOS, GMOS	NB	BACKGROUND NOISE, NOISE REDUCTION	EG 202 396-3 [10]
	2e OR 2a SIGNALS	SIMULATION	FOUR…SIX QUALITY FEATURES (COLORATION, NOISINESS, DISCONTINUITY, LOUDNESS)	NB	SAME AS P.863	P.AMD [25]
	2e OR 2a SIGNALS	SIMULATION	FOUR…SIX QUALITY FEATURES (COLORATION, NOISINESS, DISCONTINUITY, LOUDNESS)	SWB	SAME AS P.863	P.AMD [25]
MONITORING	1e SIGNAL	MEASUREMENT	MOS-LQON	NB	SAME AS P.862	P.563 [32], ANIQUE+ [1]
MAINTENANCE	2e SIGNALS	MEASUREMENT	MOS-LQOW	WB	SAME AS P.862	P.862 [39] (SEE NOTE 2 BELOW)
	2e OR 2a SIGNALS	MEASUREMENT	MOS-LQOM	NB, WB (SEE NOTE 3 BELOW), SWB	SAME AS P.863	P.863 [24]
	1e SIGNAL	MEASUREMENT	MOS-CQON	NB	SAME AS G.107	P.562 (CCI) [31]
	2e SIGNALS	MEASUREMENT	MOS-CQEM	NB, WB, SWB	SAME AS G.107	G.107 (NONINTRUSIVE)
	PARAMETERS	MEASUREMENT	MOS-LQON	NB	IP NETWORK IMPAIRMENTS ON THE ONE-WAY LISTENING QUALITY	P.564 [33]
	PARAMETERS	MEASUREMENT	MOS-LQOW	WB	IP NETWORK IMPAIRMENTS ON THE ONE-WAY LISTENING QUALITY	P.564 ANNEX B [33]
	2e SIGNALS	MEASUREMENT	MOS-CQO	NB	ECHO, SIDETONE	PESQM [2]

Type of input: 1 = one signal; 2 = two signals; a = acoustic signal or recording; e = electrical signal. Target: MOS = mean opinion score; CQ = conversational quality; LQ = listening quality; E = estimated during network planning; O = objective using signal-based measures; N = narrow-band; M = mixed narrow-, wide- and/or superwide-band; W = wide-band. Audio bandwidth: NB = 300–3,400 Hz; WB = 50–7,000 Hz; SWB = 50–14,000 Hz. Notes: 1) Full WB and SWB models covering all channel elements not yet available; 2) ITU-T Rec. P.862 and related models are no longer recommended by ITU-T SG12, and should be replaced by ITU-T Rec. P.863; and 3) no separate WB mode available; to be used in its SWB mode.

coloration, noisiness, and nonoptimum loudness, and more fine-grained features such as low- and high-frequency coloration, slow- and fast-varying time-localized distortions as well as the level and variation of background noise [41].

For network monitoring, a variety of approaches is conceivable; thus model usage should be guided based on the number of available signals and/or if one-way or two-way communication is sought. For one-way communications, (i.e., listening-only) two methods may be employed. First, if only a single electrical signal is available during network operation, reference-free models such as [32] and [1] should be used (the so-called nonintrusive measurement). Second, if two electrical or acoustically recorded signals are available, reference-based models such as [39] and [24] should be considered. Note that

to perform reference-based quality measurement, a clean reference signal needs to be injected into a piece of equipment or a call connection, thus temporarily disrupting the network or call session. For two-way communications, or conversational quality measurement, users have four options. Within a reference-based paradigm, practitioners may employ the clean and processed signals to estimate impairment factors that can be forwarded to parametric models, as is recommended by ITU-T Rec. P.562 [31]. Alternately, users may opt to use the nonintrusive version of the E-model. Third, parameters collected from packet headers can be used for parametric estimation using approaches that can be validated according to [33] (see the section "Protocol-Information-Based Models"). While the abovementioned methods fall within the parametric or hybrid paradigms, pure signal-based models may also be applied to predict conversational quality; one such approach, though not recommended by ITU, is described in [2].

QUALITY DIMENSIONS OF INTEREST CAN INCLUDE LISTENING EFFORT, ARTICULATION, NATURALNESS, CONTINUITY/FLUENCY, AND PROSODY SIMILARITY WITH NATURAL SPEECH, TO NAME A FEW.

FUTURE TRENDS AND CHALLENGES

Full-reference signal-based models were widely developed in the 1990s, and the last decade witnessed a widespread domination of the PESQ model, mostly due to its high correlation with subjective quality scores. Needless to say, speech transmission systems have evolved over the last decade, employing more complex signal processing algorithms, such as speech enhancement. As expected, PESQ and WB-PESQ performance did not follow suit, leading to the development of its successor—ITU-T Rec. P.863 (POLQA). While the new POLQA model [24] is intended to cover all relevant in-use transmission scenarios and equipment, the model is still restricted to short sentences of approximately 6–20 s, far below the duration of a typical phone call, which ranges between one and two minutes. Nevertheless, a major trend is the development of models that predict "full-length" call quality. The first steps have already been taken [3], [65], and a standard has been put forward [11], though not yet covering WB or SWB transmission. Clearly, further research is still needed before a general-purpose tool is available. One possibility is to investigate temporal-integration strategies that combine multiple short-duration P.863 estimations into a final longer-duration call quality rating. Similar considerations apply for no-reference methods including protocol-header-based and hybrid approaches. Some of these models can already account for longer observation windows but need to be adapted to WB and SWB quality prediction. Here, it will be desirable to achieve equally stable and well-validated models as P.863.

Another major trend pursued by the speech quality measurement community has been the development of reliable multidimensional quality models for enhanced speech (e.g., noise suppressed, bandwidth extension, and dereverberated speech), where unwanted perceptual artifacts, residual noise, and signal-component distortions need to be detected and quantified [16]. The same holds true for the variety of other signal processing equipment used in today's networks and terminal elements (e.g., level-adjustment and echo cancellation); such equipment can be best characterized with dimension-based approaches such as the one described in [25]. Multidimensional quality models can also play a pivotal role in characterizing human-machine communication, such as diagnosing the quality of synthesized speech and of spoken dialogue systems [15], [49]. Quality dimensions of interest can include listening effort, articulation, naturalness, continuity/fluency, and prosody similarity with natural speech, to name a few.

Atypical speech constitutes a range of novel conditions for most existing standard speech quality models. These models are mostly tested using the speech and listening of healthy adult native speakers of a few selected languages. Model performance may be poor for speech and hearing capabilities that are not represented in the test conditions.

While speech "quality" serves as an essential performance measure in typical telecommunication conversation settings, other measures provide more specifically useful information. Currently, objective measurement of speech "intelligibility" is actively researched. Commonly, "intelligibility" measures the percentage of words or subword units understandable to native speakers with healthy hearing and cognition. Speech intelligibility is a component of speech quality, in the sense that good intelligibility is necessary but not sufficient for good quality. For instance, robotic speech may be perfectly intelligible but not natural and hence not high quality. Speech intelligibility measures are particularly useful in degraded conditions, such as noisy and reverberated speech, talkers with speech impairments, and listeners with hearing impairments. For instance, Hu and Loizou [21] reported that most existing noisy-speech enhancement algorithms improve speech quality but hardly improve speech intelligibility. In some task-oriented, mission-critical applications, it might be preferable to configure speech enhancement processing to maximize speech intelligibility, perhaps at the expense of weighing less other attributes of speech quality such as naturalness. In human-machine communication wherein speech reception is by a machine, "quality" may be measured to enable maximization of machine "intelligence." While a measure of intelligibility may be appropriate for automatic speech recognition, the measure might not be adequate for speaker identification.

Finally, speech and audio quality are an integral part of the user's perception of multimedia quality. Future research directions will strive to develop models that combine audio and video quality information and which better reflect human interaction behavior. Such models may also be able to cover surround sound and three-dimensional video with

stereoscopic rendering techniques. It is also possible that future models will incorporate adaptable user-specific parameters to truly represent a user's quality of experience (QoE). A major challenge witnessed today is the lack of publicly available data to develop and test such models. Devising models and/or model design methods that have low subjective-data costs is an ongoing challenge. Crowdsourcing approaches (e.g., Amazon's Mechanical Turk) may contribute to the solution of this challenge [56].

AUTHORS

Sebastian Möller (sebastian.moeller@telekom.de) studied electrical engineering at the universities of Bochum (Germany), Orléans (France), and Bologna (Italy). He received a doctor-of-engineering degree in 1999 and the venia legendi with a book on the quality of telephone-based spoken dialogue systems in 2004. In 2005, he joined Deutsche Telekom Laboratories, TU Berlin, and in 2007, he was appointed professor for quality and usability at Technical University (TU) Berlin. His primary interests are in speech signal processing, speech technology, and quality and usability evaluation. Since 1997, he has taken part in ITU–T Study Group 12, where he is currently corapporteur of Question Q.8/12.

Wai-Yip Chan (geoffrey.chan@queensu.ca), also known as Geoffrey Chan, received his B.Eng. and M.Eng. degrees from Carleton University, Ottawa, and his Ph.D. degree from the University of California, Santa Barbara. He is currently with the Department of Electrical and Computer Engineering, Queen's University, Canada. He has held positions with the Communications Research Centre, Bell Northern Research (Nortel), McGill University, and Illinois Institute of Technology. His research interests are in multimedia signal processing and communications. He is an associate editor of *EURASIP Journal on Audio, Speech, and Music Processing*, and a member of the IEEE Signal Processing Society Speech and Language Technical Committee.

Nicolas Côté (nicote@free.fr) studied audiovisual engineering at the University of Valenciennes, France. He received a master's degree in acoustic and signal processing applied to music signals from the University of Paris VI in 2005. He joined France Télécom R&D in Lannion, France; Deutsche Telekom Laboratories, TU Berlin, Germany, in 2005; and received a doctor-of-engineering degree in 2010 for his work on the integral and diagnostic intrusive prediction of speech quality. He now works as a scientific researcher at the Université de Bretagne Occidentale, in Brest, France. His research interests include quality assessment of speech transmission and sound reproduction systems.

Tiago H. Falk (tiago.falk@ieee.org) received the B.Sc. degree from the Federal University of Pernambuco, Brazil, in 2002, and the M.Sc. and Ph.D. degrees from Queen's University, Canada, in 2005 and 2008, respectively, all in electrical engineering. Since 2010, he has been an assistant professor at the Institut National de la Recherche Scientifique (INRS-EMT) in Montreal, Canada. His research interests are in multimedia quality measurement and enhancement. His work has engendered numerous awards, including the IEEE (Kingston Section) Ph.D. Research Excellence Award, Best Student Paper Awards at ICASSP (2005) and IWAENC (2008) conferences, and the Newton Maia Young Scientist Award.

Alexander Raake (alexander.raake@telekom.de) received his doctoral degree in electrical engineering and information technology from Ruhr-University Bochum, Germany, in 2005, and his electrical engineering diploma from RWTH Aachen, Germany, in 1997. From 1998 to 1999 he was a researcher at EPFL, Switzerland. Between 2004 and 2009 he held postdoc and senior scientist positions at LIMSI-CNRS, France, and Deutsche Telekom Laboratories, Germany, respectively. Since 2009, he has been an assistant professor at Deutsche Telekom Laboratories, TU Berlin. His research interests are in multimedia technology and QoE. Since 1999, he has been active in ITU-T, currently as corapporteur for Q.14/12 on audiovisual quality.

Marcel Wältermann (marcel.waeltermann@telekom.de) studied electrical engineering at Ruhr-University Bochum, Germany. In 2005, he graduated in the area of communication acoustics. After a two-year engagement at Ruhr-University Bochum, he now works as a scientific researcher at Deutsche Telekom Laboratories, TU Berlin, on quality models for transmitted speech on the basis of perceptual dimensions. Further interests include communication acoustics and speech signal processing. He has been corapporteur for Question 8/12 of ITU–T Study Group 12 since 2009.

REFERENCES

[1] *Auditory Non-Intrusive Quality Estimation Plus (Anique+): Perceptual Model for Non-Intrusive Estimation of Narrowband Speech Quality, ATIS-PP-0100005.2006, American National Standards Institute*, 2006.

[2] R. Appel and J. G. Beerends, "On the quality of hearing one's own voice," *J. Audio Eng. Soc.*, vol. 50, no. 4, pp. 237–248, 2002.

[3] J. Berger, A. Hellenbart, R. Ullmann, B. Weiss, S. Möller, J. Gustafsson, and G. Heikkilä, "Estimation of 'quality per call' in modelled telephone conversations," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'08)*, Las Vegas, NV, 2008, pp. 4809–4812.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.

[5] S. Broom, "VoIP quality assessment: Taking account of the edge-device," *IEEE Trans. Audio, Speech Lang. Processing (Special Issue on Objective Quality Assessment of Speech and Audio)*, vol. 14, no. 6, pp. 1977–1983, Nov. 2006.

[6] A. Clark, "Modeling the effects of burst packet loss and recency on subjective voice quality", in *Proc. Internet Telephony Workshop (IPtel'01)*, New York, Apr. 2001, pp. 1–5.

[7] N. Côté, V. Koehl, V. Gautier-Turbin, A. Raake, and S. Möller "An intrusive super-wideband speech quality model: DIAL," in *Proc. 11th Annu. Conf. Int. Speech Communication Association (Interspeech'10)*, Makuhari, Japan, 2010, pp. 1317–1320.

[8] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality. Berlin*: Springer-Verlag, 2011.

[9] A. Ekman and B. Kleijn, "Improving quality prediction accuracy of P.563 for noise suppression," in *Proc. Int. Workshop for Acoustic Echo and Noise Control, CD_ROM*, 2008.

[10] "Speech processing, transmission and quality aspects (STQ); speech quality performance in the presence of background noise. Part 3: Background noise transmission—Objective test methods," *Europ. Telecomm. Standardization Institute,* Sophia Antipolis, France, *ETSI EG 202 396-3*, 2008.

[11] "Speech processing, transmission and quality aspects (STQ); estimating speech quality per call," *Europ. Telecomm. Standardization Institute,* Sophia Antipolis, France, *ETSI TR 102 506*, 2007.

[12] T. H. Falk and W.-Y. Chan, "Performance study of objective speech quality measurement for modern wireless—VoIP communications," *EURASIP J. Audio, Speech Music Processing*, vol. 2009, Article ID 104382, 11 pages.

[13] T. H. Falk and W.-Y. Chan, "Hybrid signal-and-link-parametric speech quality measurement for VoIP communications," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 16, no. 8, pp. 1579–1589, Nov. 2008.

[14] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Processing Lett.*, vol. 13, no. 2, pp. 108–111, Feb. 2006.

[15] T. H. Falk and S. Möller, "Towards signal-based instrumental quality diagnosis for text-to-speech systems," *IEEE Signal Processing Lett.*, vol. 15, pp. 781–784, 2008.

[16] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech Lang. Processing (Special Issue on Processing Reverberant Speech: Methodologies and Applications)*, vol. 18, no. 7, pp. 1766–1774, Sept. 2010.

[17] K. Genuit, "Objective evaluation of acoustic-quality based on a relative approach," in *Proc. Inter-Noise* 1996, Liverpool, England, paper 1061, pp. 1–6.

[18] J. D. Gibson, *The Communications Handbook*. Boca Raton, FL: CRC, 2002.

[19] P. Gray, M. P. Hollier, and R. E. Massara, "Non-intrusive speech quality assessment using vocal-tract models," *Proc. Inst. Elect. Eng. Vision, Image, Signal Processing*, vol. 147, no. 6, pp. 493–501, Dec. 2000.

[20] M. Guéguin, R. Le Bouquin-Jeannès, V. Gautier-Turbin, G. Faucon, and V. Barriac, "On the evaluation of the conversational speech quality in telecommunications," *EURASIP J. Adv. Signal Processing*, vol. 2008, 2008, Article ID 185248, 15 pages.

[21] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, July 2007.

[22] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, Eds., "RTP: A transport protocol for real-time applications," *IETF RFC 3550*, Internet Engineering Task Force, Freemont, CA, July 2003.

[23] T. Friedman, R. Caceres, iand A. Clark, Eds., "RTCP extended report (XR)," *IETF RFC 3611*, Internet Engineering Task Force, Freemont, CA, Nov. 2003.

[24] ITU, "Perceptual objective listening quality assessment," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863, 2011.

[25] ITU, "Draft requirement specification for P.AMD (perceptual approaches for multi-dimensional analysis)," Source: Deutsche Telekom AG, ITU-T SG12 WP2 Meeting, 17 Sept. 2010, Berlin, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Contr. COM 12-143, 2010.

[26] ITU, "Performance of the joint POLQA model," Source: Opticom, TNO, SwissQual, ITU-T SG12 WP2 Meeting, Sept. 17, 2010, Berlin, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Contr. COM 12-148, 2010.

[27] ITU, "Perceptual correlates of the E-Model's impairment factors," Source: Federal Republic of Germany (Authors: M. Wältermann and S. Möller), ITU-T SG12 Meeting, Oct. 17–21, Int. Telecomm. Union, Geneva, Switzerland, ITU-T Delayed Contribution D.071, 2005.

[28] ITU, "Handbook on telephonometry," Int. Telecomm. Union, Geneva, Switzerland, ITU-T, 1992.

[29] ITU, "The E-model: A computational model for use in transmission planning," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.107, 2009.

[30] ITU, "In-service non-intrusive measurement device—Voice service measurements," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.561, 2002.

[31] ITU, "Analysis and interpretation of INMD voice-service measurements," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.562, 2004.

[32] ITU, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.563, 2004.

[33] ITU, "Conformance testing for voice over ip transmission quality assessment models," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.564, 2007.

[34] ITU, "Methods for subjective determination of transmission quality", Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.800, 1996.

[35] ITU, "Methodology for the derivation of equipment impairment factors from instrumental models," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.834, 2002.

[36] ITU, "Extension of the methodology for the derivation of equipment impairment factors from instrumental models for wideband speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.834.1, 2009.

[37] ITU, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.835, 2003.

[38] ITU, "Objective quality measurement of telephone-band (300–3400 Hz) speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.861, 1996.

[39] ITU, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862, 2001.

[40] ITU, "Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862.2, 2005.

[41] ITU, "Multidimensional subjective testing methodology," Source: Rapporteurs of Q.7/12, ITU-T SG12 Meeting, Geneva, Switzerland, Jan. 18–27, 2011, Int. Telecomm. Union, Geneva, Switzerland, ITU-T TD 367(GEN), 2011.

[42] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. New York: Springer-Verlag, 2005.

[43] C. Jin and R. Kubichek, "Vector quantization techniques for output-based objective speech quality," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, May 1996, vol. 1, pp. 491–494.

[44] N. O. Johannesson, "The ETSI computation model: A tool for transmission planning of telephone networks," *IEEE Commun. Mag.*, vol. 35, no. 1, pp. 70–79, Jan. 1997.

[45] N. O. Johannesson, "Echo canceller performance characterized by impairment factors," *ITU-T Speech Quality Experts Group Meeting,* Ipswich, U.K., Sept. 23–27, Int. Telecomm. Union, Geneva, Switzerland, *Doc. IP–16*, 1996.

[46] D.-S. Kim and A. Tarraf, "ANIQUE+: A new American national standard for non-intrusive estimation of narrowband speech quality," *Bell Labs Tech. J.*, vol. 12, no. 1, pp. 221–236, May 2007.

[47] J. Liang and R. Kubichek, "Output-based objective speech quality," in *Proc. IEEE Vehicular Technol. Conf.,* Stockholm, Sweden, pp. 1719–1723, 1994.

[48] L. Malfait, J. Berger, and M. Kastner, "P.563—The ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[49] S. Möller, F. Hinterleitner, T. H. Falk, and T. Polzehl, "Comparison of approaches for instrumentally predicting the quality of text-to-speech systems," in *Proc. 11th Annu. Conf. Int. Speech Communication Association (Interspeech'10)*, Makuhari, Japan, Sept. 26–30, 2010, pp. 1325–1328.

[50] S. Möller, F. Kettler, H.-W. Gierlich, N. Côté, A. Raake, and M. Wältermann, "Extending the E-model to better capture terminal effects," in *Proc. 3rd Int. Workshop on Perceptual Quality of Systems (PQS'10)*, Bautzen, Germany, 2010.

[51] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Norwell, MA: Kluwer, 2000.

[52] S. R. Quackenbush, T. P. Barnwell, and M. A. Clemens, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[53] A. Raake, "Short- and long-term packet loss behavior: Towards speech quality prediction for arbitrary loss distributions," *IEEE Trans. Audio, Speech Lang. Processing (Special Issue on Objective Quality Assessment of Speech and Audio)*, vol. 14, no. 6, pp. 1957–1968, Nov. 2006.

[54] A. Raake, S. Möller, M. Wältermann, N. Côté, and J.-P. Ramirez, "Parameter-based prediction of speech quality in listening context—Towards a WB E-model," in *Proc. 2nd Int. Workshop Quality of Multimedia Experience (QoMEX'10)*, June 21–23, 2010, pp. 182–187.

[55] A. Raake, *Speech Quality of VoIP—Assessment and Prediction*. Chichester, West Sussex, U.K.: Wiley, 2006.

[56] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "CROWDMOS: An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'11)*, Prague, Czech Republic, 2011, pp. 2416–2419.

[57] D. L. Richards, *Telecommunications by Speech*. London: Butterworths, 1973.

[58] K. Scholz, "Instrumentelle Qualitätsbeurteilung von Telefonbandsprache beruhend auf Qualitätsattributen (Instrumental quality assessment of telephone-band speech based on quality attributes)," Doctoral Dissertation *(Arbeiten über Digitale Signalverarbeitung, no. 32)*. Aachen, Germany: Shaker Verlag, 2008.

[59] D. Sen, "Predicting foreground SH, SL and BNH DAM scores for multidimensional objective measure of speech quality," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, Montreal, May 2004, vol. 1, pp. 493–496.

[60] W. D. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. ICASSP'77,* Hartford, CT, 1977, pp. 204–207.

[61] M. Wältermann, A. Raake, and S. Möller, "Quality dimensions of narrowband and wideband speech transmission," *Acta Acust. United Acust.*, vol. 96, no. 6, pp. 1090–1103, 2010.

[62] M. Wältermann, K. Scholz, S. Möller, L. Huo, A. Raake, and U. Heute, "An instrumental measure for end-to-end speech transmission quality based on perceptual dimensions: Framework and realization," in *Proc. Interspeech* 2008, pp. 22–26.

[63] M. Wältermann, I. Tucker, A. Raake, and S. Möller, "Analytical assessment and distance modeling of speech transmission quality," in *Proc. 11th Ann. Conf. Int. Speech Communication Association (Interspeech'10)*, Makuhari, Japan, 2010, pp. 1313–1316.

[64] M. Wältermann, A. Raake, and S. Möller, "Extension of the E-model towards super-wideband speech transmission," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'10),* Dallas, TX, 2010, pp. 4654–4657.

[65] B. Weiss, S. Möller, A. Raake, J. Berger, and R. Ullmann, "Modeling conversational quality for time-varying transmission characteristics," *Acta Acoust. United Acoust.*, vol. 95, no. 6, pp. 1140–1151, 2008.

[SP]

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

Zhou Wang and Alan C. Bovik

Reduced- and No-Reference Image Quality Assessment

The natural scene statistic model approach

Recent years have witnessed dramatically increased interest and demand for accurate, easy-to-use, and practical image quality assessment (IQA) and video quality assessment (VQA) tools that can be used to evaluate, control, and improve the perceptual quality of multimedia content in a wide variety of practical multimedia signal acquisition, communication, and display systems. There is a vast and increasing proliferation of such content over both wireline and wireless networks. Think of the Internet: Youtube, Facebook, Google Video, Flickr and so on; networked high-definition television (HDTV), Internet Protocol TV (IPTV) and unicast home video-on-demand (Netflix and Hulu, for example); and an explosion of wireless video traffic that is expected to more than double every year over the next five years [1]. In such an environment of extreme growth, limited bandwidths, and diverse content, resolutions, and quality, there is considerable concern regarding how the quality of service (QoS) of videos being delivered can be managed. In short, there is currently no practical method for accurately monitoring the perceptual quality of this vast proliferation of video data.

A number of successful algorithms have been created that can predict subjective visual quality of a distorted image or video signal—in agreement with human opinions of visual quality—when a (presumed) "pristine" signal is fully available [2].

Digital Object Identifier 10.1109/MSP.2011.942471
Date of publication: 1 November 2011

Yet, in most present and emerging practical real-world visual communication environments, such full-reference (FR) methods are not useful since the reference signals are not accessible at the receiver side (or perhaps at all). What are really needed are I/VQA algorithms that can operate with little or no reference signal information at all; in other words, by operating on the visual signal of interest directly, rather than by extensive comparison. Indeed, the assumption of a supposedly "pristine" reference image or video is highly suspect; even under the most ideal and controlled circumstances, a captured optical signal will inevitably suffer from some kind of distortion [3].

CREATING AUTONOMOUS ALGORITHMS THAT DEPEND ON MUCH LESS SPECIFIC INFORMATION FROM ANY REFERENCE SIGNAL IS NOW AN INTENSE FOCUS OF RESEARCH.

Thus, creating autonomous algorithms that depend on much less specific information from any reference signal is now an intense focus of research. Such algorithms fall into two categories: reduced-reference (RR) and no-reference (NR) (or blind) IQA and VQA algorithms. In the former category, a reference signal is assumed only partially accessible (in the form of selected features); typically, the amount of data from the reference signal is significantly less than in the reference signal itself. In the latter category, reference signal information is deemed completely inaccessible [3]. Although RR and NR algorithms (that accord with perception) have been desired for a very long time, progress has been slow. Yet the need is large since today's video consumers have become increasingly savvy about the capabilities of digital video, and expectations regarding the QoS of delivered visual multimedia have risen significantly.

If such RR and/or NR visual QA algorithms could be created, they could be deployed as agents over wide-area data communications and visual surveillance networks by embedding them in smart routers, set-top boxes, smart phones, cameras, tablets and laptops. They could be used as primary QoS tools that could feed back time-varying visual signal quality information, enabling source adaptation and distributed network control mechanisms to adapt resource allocation, source and channel coding, and other network parameters. In today's increasing video-centric consumer data communications environment, we think that such algorithms could represent a sea change in visual multimedia data delivery.

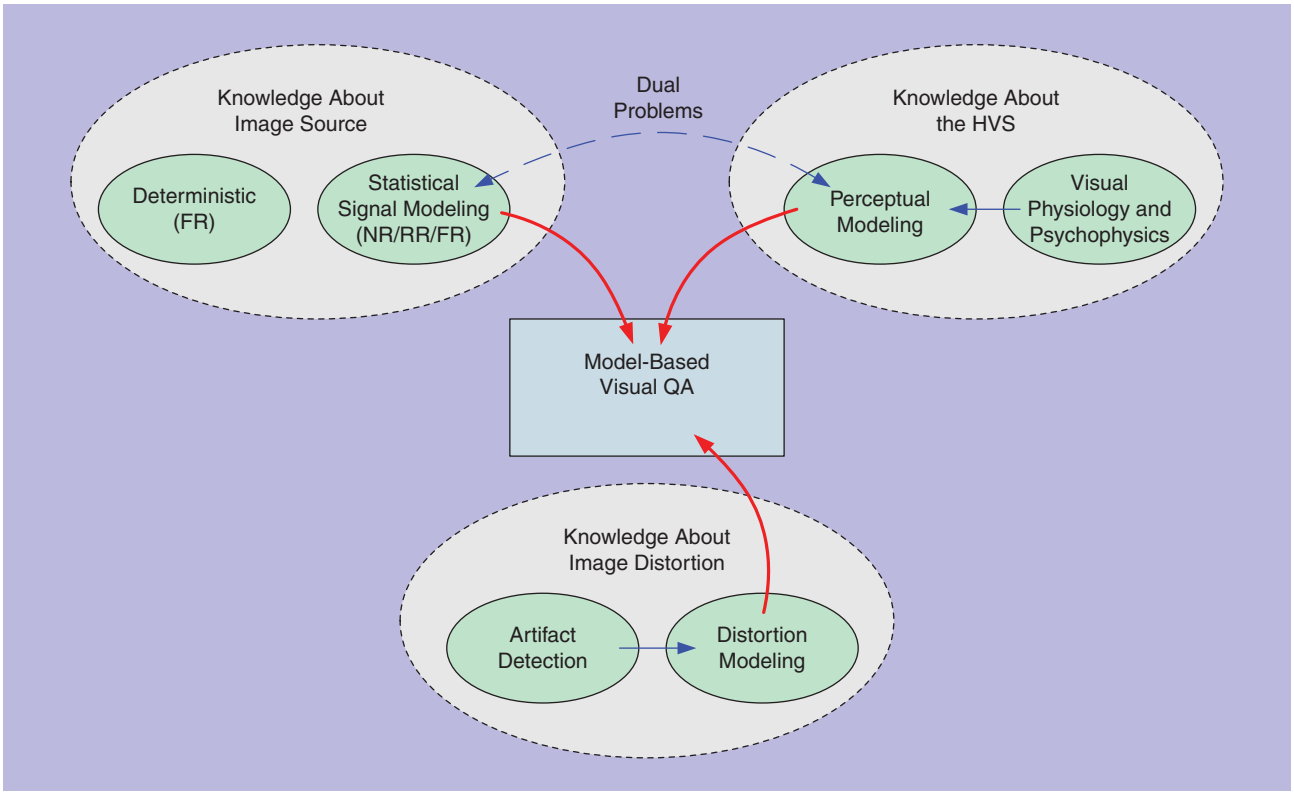
To date, a wide variety of inventive methods have been deployed toward solving the RR and NR I/VQA problems, and the universe of ideas are quite large. In attempting to describe and help the readers to understand the field, we are confronted also with the fact that the various approaches attempt to solve diverse problems that operate under different assumptions. It is our goal in this tutorial to clarify the issues to be solved and how they might be pragmatically approached, without clouding things by attempting a broad survey of the field. In doing so, we take a certain viewpoint regarding how things ought to be done.

TOWARD MODEL-BASED VISUAL QA

To better cast the foregoing discussion against the "big picture," Figure 1 depicts a "knowledge map" that contains what we regard as the essential building blocks in the design of successful visual QA models. We think that accurate modeling is the key to successful I/VQA algorithm design. Essentially, three types of knowledge may be exploited to build such models. The first is knowledge about the image source that captures the essence of what a signal ought to look like when not distorted. This can be either deterministic when the reference signal is fully accessible (FR case) or statistical when certain statistical models are available to regulate the undistorted images. The second is knowledge about image distortion, which may help detect particular artifacts created by specific distortion processes (e.g., blocking artifacts generated in JPEG compression). Furthermore, mathematical models of how distortions change image content (by degradation process, artifacts, or loss of statistical naturalness) can be used to predict distortion severity in an observed visual signal. Finally, since most applications direct visual signals toward human viewers, the third type of knowledge is about the HVS, which is based on perceptual models originated from visual physiology and psychophysical studies. We believe that models derived from these three types of knowledge should not be disjoint. In particular, statistical signal modeling and perceptual modeling may be understood as dual problems, as discussed later. By analogy, communication systems embody knowledge of transmitter (analogously visual signal model), channel (distortion model), and receiver (perceptual model). The more information that is available regarding transmitter, channel, and receiver, the better job of communication that can be accomplished. The most successful design will use and combine models of all three. Most visual QA algorithms can be understood using such a unified modeling framework, even if not expressed in such terms.

NATURAL SCENE STATISTIC MODELS

We believe that there is a category of statistical models that comes close to embodying the three-fold modeling objectives just described, and that provides the most promising basis for successful RR and NR QA algorithm design. As we shall see, these so-called natural scene statistic (NSS) models are highly attractive in a number of ways: they reliably capture low-level statistical properties of images (hence are very general and flexible models); they can be used to measure the destruction of "naturalness" introduced by distortions (enabling effective distortion models); and they accurately describe the statistics to which the visual apparatus has adapted and evolved over the millennia (and so, are regarded as direct duals of low-level perceptual models). A "natural scene" is one captured by an optical camera, and can include both naturalistic (e.g., trees and grass) content as well as man-made indoor and outdoor scenery. The term is meant to distinguish "natural" from artificial image



[FIG1] **Knowledge map expressing the elements required to construct successful visual QA models.**

creation processes such as computer graphics. We will describe specific NSS models that appear to be well suited for RR and NR QA algorithm design. We will also point out ways in which NSS models can be improved for QA applications, e.g., by incorporating perceptual information into them.

NSS models seek to capture the natural statistical behavior of images, rather than assuming deterministic knowledge of the image source (as in FR QA). Such prior models of image statistics enables the use of a rich groundwork of Bayesian statistical methods, and are rooted in the widely accepted view of biological perceptual systems in computational neuroscience and psychophysics, that the visual apparatus is highly adapted to the natural environment, and has evolved to most efficiently extract visual information from it [4], [5].

What constitutes a useful model of natural image or video statistics? The most important criteria is that the model be regular, in the sense that any natural image that has not been distorted by unnatural distortions can be expected to follow the model with a high degree of confidence. In recent years, a number of NSS models have been derived for still images that are highly regular. Common NSS models used for both modeling of perception and for image processing applications have been developed around theories of sparse coding by the visual brain [4], [18], [19] and by the observed (self-similar) scaling properties of natural images [28], [29]. Importantly, the visual brain appears to have both to have evolved to "match" the statistics of natural images [19] and to seek a efficient, decorrelated representations of image information, as evidenced by the fact that the principal components (or independent components) of natural images closely resemble the spatial responses of cortical neurons [61].

While it is beyond the scope of this article to describe the function and coding processes in visual cortex, or the broader spectrum of NSS models that have been proposed (a good introduction can be found in [62]), the connections between NSS and perceptual processes are important, since the ultimate goal of objective visual QA algorithms is to predict human behavioral responses when evaluating visual quality. Of particular importance in this context are NSS models that are sensitive to image "unnaturalness" introduced by distortion. The NSS models that we describe later on are quite useful in this regard, as they can be used to successfully predict the type and degree of perceptual quality loss introduced by common distortions.

Of course, the QA of moving pictures, or videos, can be accomplished in a limited manner by applying still-image NSS models on a frame-by-frame basis. More desirable, however, would be NSS models that capture the statistics of naturalistic videos in a natural manner. There is evidence that the spatio-temporal vision system has adapted to the natural statistics of moving images to achieve efficient encoding of the large volume of data [63]. While some progress has been made on characterizing the statistics of optical flow fields under rather rigid

assumptions [64], there does not yet exist any established statistically regular model of the natural spatiotemporal statistics of video data. The problem is greatly complicated by the complexity of the natural motions of objects, and of the sensor. As such, NSS-based VQA remains very much an open problem. Because of this, the design of (FR) VQA algorithms that are not distortion-specific has been largely driven by structural or perceptual models [35], [55], [65], [66].

THE IDEA OF RR QA WAS FIRST CONCEIVED IN THE 1990s AS A PRAGMATIC APPROACH TO REAL-TIME VIDEO QUALITY MONITORING OVER MULTIMEDIA COMMUNICATION NETWORKS.

STATISTICAL IMAGE QA

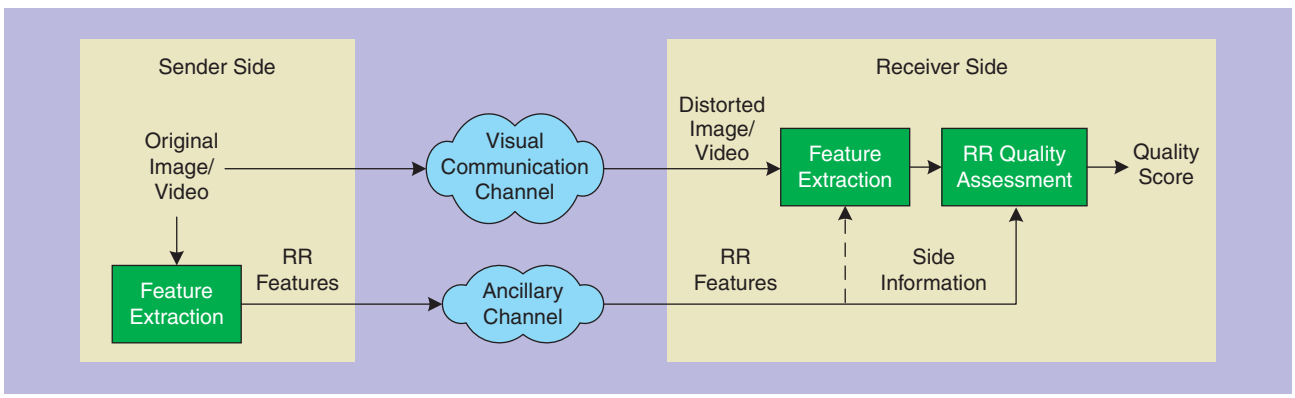
In essence, NSS-based IQA algorithms seek to capture statistical regularities of natural images and to quantify how these regularities are modified or lost when distortions occur. Since these methods do not necessarily rely on directly detecting or quantifying specific image artifacts, such algorithms have the potential to be more widely applicable than distortion-specific approaches. Such "holistic" NR QA algorithms could operate by measuring a "distance" from naturalness, thereby gauging how severely distorted a visual signal is by how "far" it lies from the space of natural images. However, it is also possible to distinguish distortions by type according to the "direction" in which the distorted signal lies away from "NSS space."

The first successful NSS model-based IQA algorithm was the FR Visual Information Fidelity (VIF) index [6], which uses a wavelet-domain Gaussian scale mixture (GSM) model (as described later in the context of RR algorithms) [7] that captures both the marginal distributions of wavelet coefficients and the magnitude dependencies of neighboring coefficients across space, scale, and orientation. This algorithm has exhibited excellent performance in two large human studies [8], [9]. The GSM was also used to modify the FR multiscale structural similarity (SSIM) index [10] by weighting local information content. The resulting information-weighted SSIM (IW-SSIM) index delivers superior performance relative to the state of the art against human subjectivity, as shown on multiple public databases [11]. These results further motivate the use of NSS for the design of visual QA algorithms. How then, to use these models when the amount of information from the reference image is reduced or eliminated? As we will show, promising results have been achieved using NSS for still image RR and NR QA, for both distortion-specific and "holistic" problems.

TOWARD RR IMAGE QA

The idea of RR QA was first conceived in the 1990s [12] as a pragmatic approach to real-time video quality monitoring over multimedia communication networks. Figure 2 depicts the general idea of how an RR image or video QA system works [3]. At the sender side, a feature extractor is applied to the reference visual signal. The extracted features are transmitted to the receiver as side information through an ancillary channel. It is usually assumed that the ancillary channel is error free. When the distorted signal is transmitted to the receiver via an error-prone channel, a feature extractor is also applied at the receiver side. This could be the same process as at the sender, or it might be adapted according to the received side information. In the final QA stage, the RR features extracted from both reference and distorted visual signals are used to compute an overall score indicating the quality of the distorted signal.

A good RR approach must achieve a good balance between the accuracy of the quality predictions and the RR data rate. One would expect that accuracy would improve monotonically as more information about the reference image is made available [3]. RR algorithms lie within two extremes: if the data rate is high enough to deliver the reference signal as side information, then an FR method can be applied at the receiver side. Conversely, if the data rate is zero (i.e., no reference side information), then an NR method is required. In practice, a maximum RR data rate is specified, which is usually quite low, since bandwidth for reference side information is effectively "stolen," since it could be used to improve the quality of the transmitted



[FIG2] General framework of an RR image or image QA system.

visual signal. This limited data rate makes the design of RR algorithms a challenging task. It puts strong constraints on the selection of RR features, which constitute the most critical component of RR algorithms. A good set of RR features should

- efficiently summarize the content of the reference visual signal
- be sensitive to specific distortions or (if of the holistic variety) be sensitive to a broad spectrum of image distortion types
- embed aspects of the signal that are perceptually relevant.

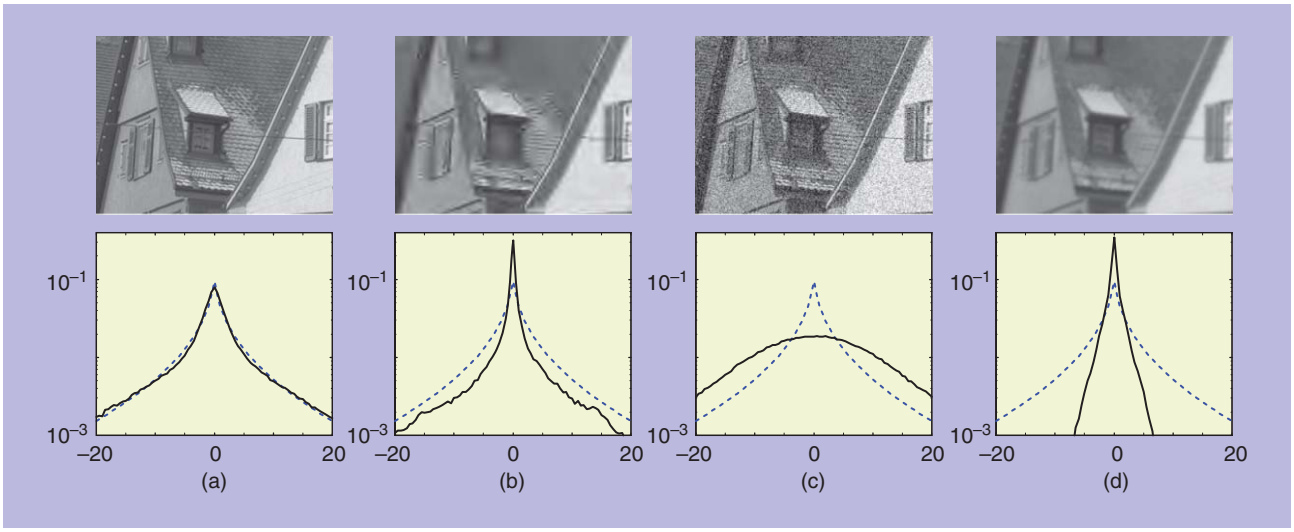
The significance of these properties can be demonstrated by a naïve example: At the sender side, randomly select image pixels (say, 1% of them) as RR features. When (side) transmitted to the receiver, they are compared on a pixel-wise basis with those in the received signal, so that the mean-squared error (MSE) or peak-signal-to-noise ratio (PSNR) between reference and received signals can be estimated. This approach is weak in several regards. First, it is difficult to keep the RR data rate low—even 1% of the pixels in a 512×512, 8 b/pixel image requires transmitting 20,976 b. An additional 47,196 b are required if the positions of the randomly selected pixels are also transmitted. This is a heavy burden—much greater than the NSS-based RR methods that we will discuss later. Second, the RR features sparsely sample the reference and do not adequately summarize the image. Third, some distortions may change only pixels not selected as RR features, and thus not be easily detected. Finally, the MSE and PSNR have poor correlations relative to the perception of visual quality [3], [8], [13], [14].

The problems exhibited in the above naïve example are instructive. Clearly, RR features should more efficiently summarize image information content, be more sensitive to image distortions, and have stronger perceptual relevance. NSS modeling provides a powerful means to approach these goals. To demonstrate this, we use an RR algorithm proposed in [15] and [60], where an NSS model of the marginal distribution of the image wavelet coefficients is employed. Owing to space limitations, we must assume that the reader is conversant with the discrete wavelet transform; otherwise an excellent easy tutorial and a deep treatment can be found in [22] and [23], respectively.

The choice of wavelet space for statistical modeling of images has a number of important underpinnings. Images are naturally multiscale and the early visual system decomposes image information in a multiscale manner, thereby representing visual signals simultaneously in localized space, frequency, and orientation [16], [17]. More importantly, natural images exhibit statistical regularities in wavelet space [5]. For example, it has been observed that the marginal distributions of natural image wavelet coefficients consistently have sharp peaks near zero and longer tails than Gaussian. This reflects specific intuitive properties of images of the real world: most of the world (and images of it) is smooth (hence many near-zero wavelet or bandpass responses). This smoothness is broken up by sparse, often large amplitude discontinuities (hence relatively many large bandpass responses). Such highly kurtotic distributions have important implications with respect to the sensory neural coding of natural scenes [18], and are a central focus of recent theories on the evolution and function of biological vision systems [19].

Figure 3 shows the histograms of the wavelet coefficients from a subband of a natural image and several distorted versions of it. A discovery in the literature of NSS is that the marginal distribution of the wavelet coefficients of natural images can be consistently well fitted by a two-parameter generalized Gaussian density (GGD) model with high accuracy [20]

$$p_m(x) = \frac{\beta}{2\,\alpha\,\Gamma(1/\beta)} \exp[-(|x|/\alpha)^{\beta}], \tag{1}$$



[FIG3] Wavelet coefficient histograms (solid curves) of (a) original "buildings" image; (b) compressed by JPEG2000; (c) with additive white Gaussian noise; and (d) blurred by a linear Gaussian kernel. The histogram in (a) is well fitted by a generalized GGD model (dashed curves). The shapes of the histograms (and the GGD fits) change in different ways for different types of distortions.

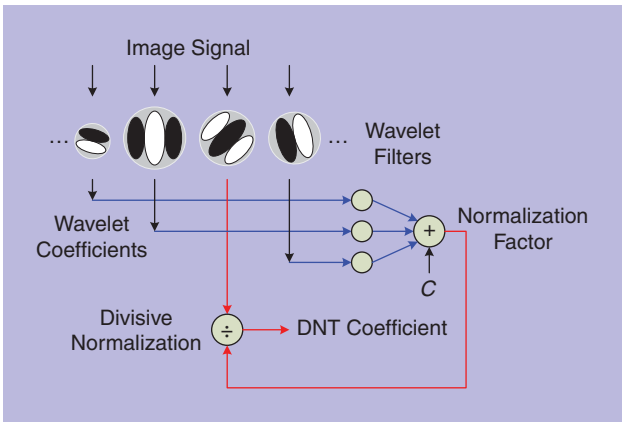
where $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ (for $a > 0$) is the Gamma function. Figure 3(a) also depicts GGD fit of the histogram of the natural images (dashed curves), which very closely approximates the true distribution (solid curve). This is important since only two parameters $\{\alpha, \beta\}$ are required to summarize the reference image coefficient histograms.

Evident from Figure 3(b)–(d) is the fact that the marginal distributions of the wavelet coefficients can change in diverse ways as the image undergoes different distortions. This is ideal for RR QA, since departures from the reference distributions (characterized by $\{\alpha, \beta\}$) can be used as a common measure to quantify the degree of distortions. One such measure is the Kullback-Leibler divergence (KLD) [21] between the model (1) and the marginal distribution of the distorted signal $q(x)$, viz., between the solid and dashed curves in Figure 3(b)–(d)

$$d(p_m \| q) = \int p_m(x) \log \frac{p_m(x)}{q(x)} dx. \tag{2}$$

This approach was used in [15] to create a holistic algorithm that achieves competitive performance relative to the full reference PSNR, using a side channel RR data rate of only 162 b/image. This is powerful evidence of the efficacy of this NSS model for IQA.

A number of earlier statistical approaches operated without any models, by extracting local sample statistics from the image. For example, in [22], simple local statistical descriptors of spatiotemporal edge and orientation activity were extracted from video sequences to form an RR video QA index for monitoring perceptual degradations in visual communication systems. In [23], local harmonic magnitudes were extracted from local image patches containing edges to create a harmonic activity map. The authors showed that different types of distortions (blocking and blurring) alter these maps in different ways, thereby providing a way to evaluate image quality with reduced reference. In [24], perceptually motivated "structural information features" (orientation, length, width, and contrast) were extracted near predicted visual fixation points. These RR features were transformed to be invariant to zoom, translation, and rotation, and stored in a database of "visual image memory." It shows good performance when predicting the quality of images compressed by JPEG and JPEG2000. In [25], the authors attempt to model the visual pathway from "front end" to cortex (starting from display, through the eyes, and ending in visual cortex), producing a set of local statistics that provide a "reduced description" of an image.



[FIG4] A simple divisive normalization scheme. A wavelet response (red output) is normalized by the summed responses of a cluster of wavelets neighboring in space, scale, and frequency. The constant *C* accounts for the saturation effect and stabilizes the DNT when the neighbor responses are low in energy. This model accords closely with neurological and psychophysical evidence, nicely explains the perceptual masking effect, and improves visual QA algorithms.

We believe it is useful to design perceptually motivated approaches using NSS-based frameworks, making it thereby possible to use Bayesian methods to achieve statistically and perceptually optimized QA. Such a stratagem is taken in [26], where NSS and perceptual models are combined in just such a manner, as described next.

Our goal here is not to give a tutorial on current models of neural processing, but there are certain aspects that require explanation in order that their relevance to QA be understood. As mentioned earlier, neurons in visual cortex effectively perform frequency- and orientation-selective waveletlike decompositions of visual data arriving from the two eyes. A nonlinear neural mechanism that has been observed is adaptive gain control (AGC), whereby each neuron's response (or in our case, wavelet coefficient response) is divided by the energy of a cluster of neighboring neuronal responses (neighboring wavelet coefficients in space, scale, and orientation) [27], as depicted in Figure 4. Such a divisive normalization transform (DNT) of the neuronal (wavelet) responses has been shown to significantly reduce statistical dependencies between the responses (coefficients) which can lead to efficient representation. Further, the statistics of the normalized coefficients have been shown to closely follow Gaussian marginal distributions [28]. More importantly, this DNT deeply affects the degree of visibility of image distortions.

The DNT is a critical perceptual model that is both useful for IQA algorithm design and that melds seamlessly with the Gaussian scale mixture (GSM) model mentioned earlier in the context of the VIF index. In fact, VIF uses a form of divisive normalization [6].

We describe the GSM model next. Suppose that $\mathbf{y}$ is a vector of wavelet responses that are locally clustered over neighboring space, scales and/or orientations. Then the GSM model is given by $\mathbf{y} = z\mathbf{u}$, where $\mathbf{u}$ is a multidimensional zero-mean Gaussian random vector, and z is a scalar random variable called a mixing multiplier. If we assume that z takes a fixed value for each selected cluster of wavelet coefficients, then putting all z values constitutes a variance field. If an accurate estimate $\hat{z}$ of z can be found for

each coefficient cluster, then dividing the observed vector of coefficients by $\hat{z}$, which accomplishes the DNT, produces a random vector that is Gaussian. This is a form of conditioning of $\mathbf{y}$ given knowledge of the variance field.

In [26], the authors used a maximum likelihood procedure [29] to estimate z and observed that the distribution of the DNT coefficients (or conditioned wavelet coefficients undergone DNT) of natural images are Gaussian, but changes in different ways in images altered with different types of distortions. Based on this observation, they form a DNT-based RR IQA algorithm, which computes the KLD (2) between the DNT coefficient histogram in the distorted image versus the best Gaussian fit to the DNT coefficients of the reference image. In their RR implementation, four features are computed from each wavelet subband: the above KLD, and the variance, kurtosis, and skewness of the DNT coefficients from that band. Their wavelet decomposition, which is a steerable pyramid [30], is taken over three scales and four orientations, yielding just 48 pieces of RR information. Yet the algorithm does quite well as measured against human subjectivity, matching the performance of the widely used FR index PSNR.

The DNT is relevant to a wide variety of neuroscience, perceptual, engineering, and in particular, QA issues. Since the DNT reduces the dependencies between the wavelet coefficients (or neural responses) over local space-scale-orientation regions, it supports the efficient coding hypothesis of early biological vision, wherein as much redundancy in representation is eliminated from low-level visual information before higher-level processes act upon it [4]. It also serves as a model of AGC, which serves to limit the dynamic range of retinal signals. AGC or DNT has an important perceptual byproduct that is easily observed: visual masking.

Visual masking is a process whereby one element of a visual signal reduces the visibility of another [31], typically of similar characteristics such as frequency or orientation. Figure 5 is an easy-to-see example, where the image of the woman is distorted everywhere by the same level of additive Gaussian noise. Although the noise statistics are unchanged across the image, it is only highly visible on the smooth regions (e.g., face), and much less visible on the more "textured" hair and scarf, and nearly imperceptible on the wicker chair backing. This effect occurs with other distortions that introduce artificial high frequencies, such as JPEG compression [32]. The significance of masking on the obscuration of spatial image distortions was observed by Girod [33] as well as Teo and Heeger [34], who proposed masking models not dissimilar to the DNT outlined above. The most successful FR IQA and VQA algorithms, such as SSIM [10] and its derivations [14], VIF [6], and Motion-Based Video Integrity Evaluator (MOVIE) [35], and the DNT-based RR algorithm outlined above, all embed masking mechanisms implemented by some kind of AGC in wavelet or scale-space.

THE DNT IS RELEVANT TO A WIDE VARIETY OF NEUROSCIENCE, PERCEPTUAL, ENGINEERING, AND IN PARTICULAR, QA ISSUES.

Naturally, NR image QA algorithms should also benefit by masking models, either by some type of DNT, or by conditioning on the signal content, or by adapting to the content in some other manner, perhaps through a training procedure. The difficulty arises since masking data is not available from a reference signal.

TOWARD NR IMAGE QA

The NR (blind) QA problem is both tantalizingly important as well as technically difficult. Yet "NR" human judgments of image quality occur with little effort. Our visual systems easily distinguish high-quality against low-quality images, and "know" what is right and wrong about them, without seeing an "original." Moreover, humans tend to agree with each other to a rather high extent. What is the mystery behind perceptual judgments of quality? Do humans have an innate ability to judge the quality of pictures relative to an unseen high standard of quality?

The answer must be positive in the sense of adaptation. Just as people from nontechnological cultures without photography interpret pictures differently from those exposed to it all their lives [36], the customers targeted by purveyors of cable, satellite, and wireless video are quite "picture savvy" with high expectations regarding the picture quality they pay for. These customers, who have observed electronic images all their lives, have collectively adapted to high-quality visual signals and to the distortions that occur. Our neural plasticity extends not only over the eons of evolution (wherein the visual systems are exposed to a large variety of natural scenes), but also over shorter spans within our lifetimes. Short-term plasticity forms the basis for our abilities of visual recognition and

[FIG5] Easy-to-see example of visual masking of white Gaussian noise by high-frequency image content.

visual memory [37], and no doubt, affects our ability to perceive a loss of quality. In other words, there are models of high-quality "reference signals" in our brains, and a learned ability to use these models to assess picture quality. High-definition-equipped readers might try to recall viewing analog TV, and how their satisfaction with respect to this older visual experience has changed.

While we do not know the exact nature of these models, clues are available from prior work on FR and RR QA. We believe that the "prior model," on which the brain relies as it perceives levels of picture quality, must be statistical and reflect the statistics of natural scenes. In this regard, the kind of NSS models we have been discussing are likely well suited for adaptation into theories of visual quality perception, and hence NR QA algorithm design.

Prior work on NR IQA algorithm design has not emphasized statistical image modeling, and most approaches presume that the distortion affecting the visual signal is known. This methods typically estimating image blur (e.g., via edge loss) [38], [39] or JPEG and JPEG2000 compression artifacts by looking for artifact signatures in spatial or spectral domains [40]–[44]. Perceptual modeling remains underutilized, although the authors of [45] utilize a psychometric model derived from subjective tests to create a perceptual blur index, which interestingly attempts to estimate blur relative to image content. Another interesting approach to blur assessment is taken in [46], where a loss of local phase coherence in the complex wavelet domain quantifies departures from image "naturalness." One NSS model-based distortion-specific NR IQA algorithm uses a GSM model to characterize correlations between the wavelet coefficients of images over scales [47]. By measuring reductions in these correlations induced by distortion, good quality prediction performance was demonstrated on JPEG2000 compressed images.

All of the above-described prior work has been geared toward still images only. The field of NR video QA has seen less progress, and almost no work at all using statistical image models. As mentioned earlier, a primary reason for this is a dearth of regular statistical models of naturalistic videos. Many proposed algorithms measure blockiness in compressed videos, e.g., by MPEG-2 [48], [49] or by H.264 [50], [51]. A usual technique is to evaluate edge-strength at block boundaries then relate it to quality. One recent NR QA method for assessing H.264-compressed video quality does use an NSS image model (Laplace/Cauchy) of the transform coefficient distributions, along with a perceptual model of contrast sensitivity and eye movement. They report good performance using their own database of videos and subjective scores [52].

Only a small amount of work has been done on the extremely difficult problem of designing NR QA algorithms that are not fixed to a single type or source of distortion. One interesting approach taken in [53] observes that the statistics of natural images tend to be locally isotropic. The authors hypothesize that image distortions destroy this property, making it possible to detect distortions. They develop an algorithm to measure the degree of local anisotropy across the image using a form of (Renyi) entropy, which is then mapped to quality scores. We implemented and tested this algorithm on the Laboratory for Image and Video Engineering (LIVE) Image Quality database [67], where it achieved a poor correlation score relative to human subjectivity; however, the idea is sound and likely could be improved by an underlying NSS model and additional features. Indeed, inspired by this, we created a simple distortion-agnostic IQA algorithm called Blind Image Integrity Notator Using DCT Statistics (BLIINDS) that uses four simple DCT-domain sample statistics computed from local windows. Inspired by the work in [53], BLIINDS-I (as we will refer it, to distinguish it from a much more evolved version of the general idea) uses two local DCT-domain entropy features and two other simple DCT statistics (kurtosis and contrast) which were to fit to half of the (content-divided) LIVE IQA Database and tested on the other half, using a simple probabilistic prediction model. The method has the virtues of conceptual and computational simplicity and achieved prediction-performance parity with the FR PSNR metric [54]. BLIINDS-I does not rely on a statistical image model, however. Instead, it uses intuitive DCT-domains sample statistics.

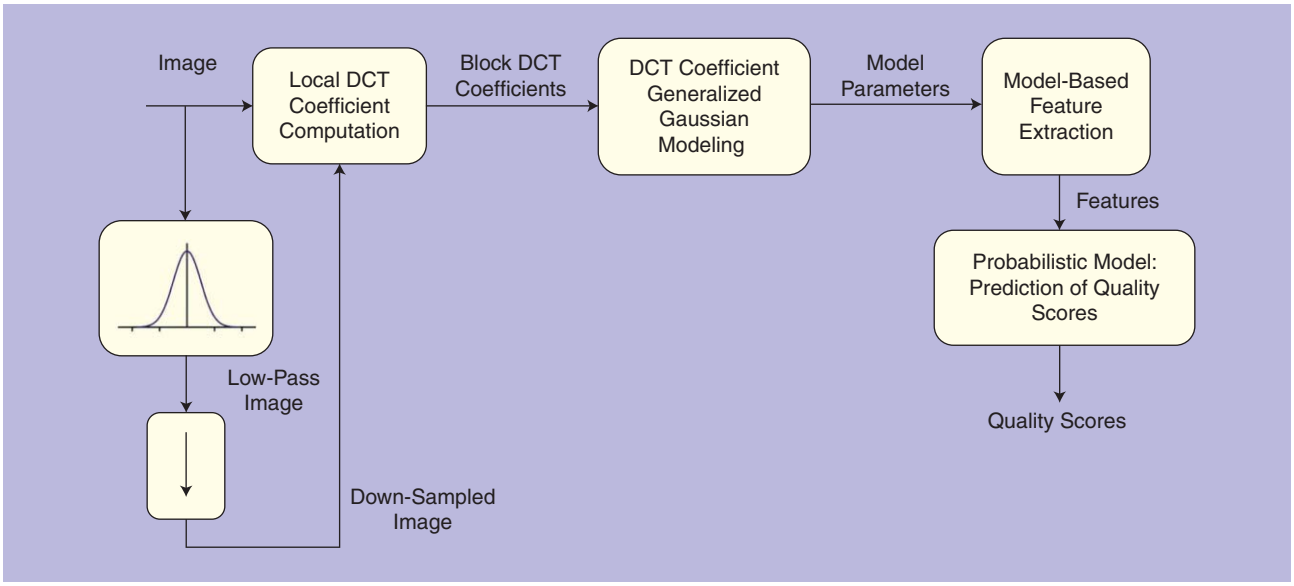
NSS MODEL-BASED APPROACHES TO NR IQA

Motivated by recent developments in NSS-based image modeling and NSS-based RR algorithm design, we have developed new NSS model-based approaches to the NR IQA problem. One exemplary method, named BLIINDS-II, retains the moniker since it still uses easily computed statistics of block DCT coefficients as features, trained on subjective scores [56]. However, BLIINDS-II is quite different from BLIINDS-I: it is model based, and uses very different features that are NSS model based.

Figure 6 diagrams the overall flow of the BLIINDS-II NR IQA algorithm. It is an instructive example, since the features used are simple and naturally defined on the NSS model.

It operates over three scales (performance has been found to remain constant if more scales are added). At the finest scale, nonoverlapping 5×5 image blocks are DCT-transformed and the resulting (non-DC) coefficients used to define statistical model-based features. The coarser scales are obtained by downsampling with a 3×3 Gaussian anti-aliasing kernel. The multiscale DCT basis is used to balance the need to approximate the natural waveletlike multiscale representation of images in the brain with the need for computational efficiency and compatibility with existing DCT-based image processing algorithms. Fortunately, a simple NSS model applies with excellent regularity to the local DCT data.

The BLIINDS-II features are also simply defined. The essential NSS model that is used is the GGD model given in (1) that has been successfully used in RR algorithm design. Specifically, the non-DCT coefficients are modeled as GGD by fitting each



[FIG6] **Flow diagram of BLIINDS-II NR IQA algorithm.**

block DCT histogram with the best-fitting GGD function. Each block is also divided into subblocks (Figure 7(a) and (b), respectively) designed to capture the radial frequency and orientation behavior, and the histogram fit is done on each of these subblocks as well. In this way, the estimated NSS model parameters are used to create all features used in BLIINDS-II.

Only very simple parametric features are extracted from the fit to the GGD NSS model: the GGD shape parameter b (sensitive to distortion signatures); coefficient of variation (CoV) z 5 $\sigma_{|X|}/\mu_{|X|}$ of the magnitudes of the GGD variates X (a normalized energy measure useful for assessing the amount of local image energy, which also accounts for masking); the ratios of energy between the radial frequency bands shown in Figure 7(a)

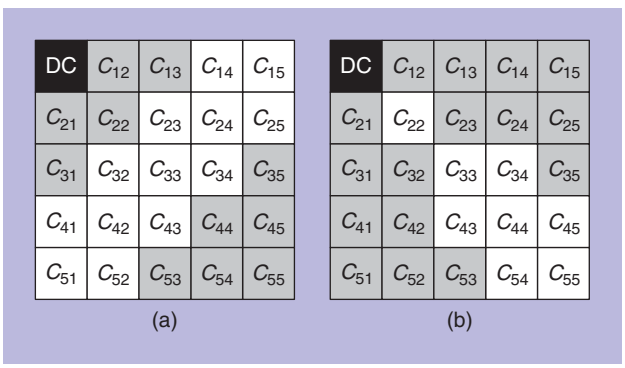
$$R_n = \frac{|E_n - \frac{1}{n-1}\sum_{j<n} E_j|}{E_n + \frac{1}{n-1}\sum_{j<n} E_j}, \tag{3}$$

where the subband energy is the band variance (the ratios indicate the relative proportions of high-, mid-, and low-frequency content in each block). Indexing the subbands outward from DC, two energy ratios are used as features: R_2 and R_3. Finally, two orientation features computed from the DCT block subdivisions shown in Figure 7(b) are used: the standard deviations of the CoV and of the shape parameters: σ_ζ and σ_β, which effectively capture changes in the orientation statistics over the three bands shown (for each scale).

Thus, just a few (six) NSS-based features are extracted using the model fit to the local DCT data. These features are computed over three scales. The features are then pooled in two ways: 1) the block feature values are averaged over the image (standard mean pooling); and (2) only the upper or lower 10% of the block feature values are averaged (depending on the trend of feature against quality). This percentile pooling strategy exploits a previously behavioral observation that the worst distortions in an image affect subjective judgments most strongly [55], [57].

BLIINDS-II is a holistic NR IQA algorithm in the sense that no specific distortion model is used to guide the design of the algorithm, and moreover, it is intended to be useful. In the absence of any distortion model, it is necessary to train NSS model based IQA algorithms using a sufficiently large and diverse database of distorted images with associated subjective data, in the form of mean opinion scores (MOS) or difference of MOS (DMOS) [67], [68].

The method of training and testing is important since any "large-scale" image database still represents a very sparse sampling of the space of all possible images. Training with such a small number of samples (relative to the image space) can lead to over fitting. Thus NR QA algorithms that are designed using train-test procedures should



[FIG7] **Block DCT coefficients divided into bands. (a) Radial frequency bands; (b) orientation bands.**

follow a few simple "rules of conduct." First, there exists a standardized protocol for obtaining the human scores that compose the subjective portion of the database [69]. Second, training and testing should be cross-validated on multiple randomized divisions of the database. We suggest a minimum of 1,000 train-test sequences over which average/median performance and standard error are taken. Lastly, each train-test sequence should randomly divide the database by content, so that content is not learned and used by the algorithm.

THE METHOD OF TRAINING AND TESTING IS IMPORTANT SINCE ANY "LARGE-SCALE" IMAGE DATABASE STILL REPRESENTS A VERY SPARSE SAMPLING OF THE SPACE OF ALL POSSIBLE IMAGES.

BLIINDS-II was trained using the above cross-validation procedure on the LIVE IQA database: 1,000 randomized train-test divisions, using 80% of the content (and distorted versions) for training, and the other 20% for testing. Training was accomplished in a simple manner: the mean and covariance of the algorithm scores and the subjective scores were fit to a multivariate Gaussian distribution to form a probability model. In the test phase, prediction is accomplished by maximizing the conditional likelihood of the subjective scores, given the observed features. Despite the simplicity of the model, the features, and the training method, the performance of BLIINDS-II is remarkably good. Over the 1,000 sequences, the Spearman rank order correlation coefficient (SROCC) and linear correlation coefficient were computed, yielding nearly equal median values of 0.91 against subjectivity, soundly beating the FR PSNR and matching the established performance of the FR SSIM index. Table 1 shows SROCC scores of BLIINDS-II against several leading FR IQA algorithms. Although the performance of BLIINDS-II does not quite match that of the best-performing FR algorithms such as multiscale SSIM (MS-SSIM) or VIF, the level of performance attained by BLIINDS-II is remarkably close to that achieved by the best algorithms that have available the reference image for comparison.

A completely different and specialized approach that can be taken is to attempt distortion identification followed by QA. Such a *two-stage* approach is taken in [58] and [59], where GSM and GGD NSS models in the wavelet-domain are used to create a holistic NR IQA algorithm with very consistent performance comparable to BLIINDS-II. The method is complementary to BLIINDS, since it seeks to determine what distortion(s) afflict an image by computing likelihoods that each distortion is present; these are used to weight multiple distortion-specific QA algorithm scores derived from the same NSS models. When trained using 1,000 iterations of cross-validation on the LIVE IQA database, very good performance is also attained, equivalent to both BLIINDS-II and the SSIM index (Table 1). The reader is referred to [59], since the Distortion Identification-Based Image Verity and Integrity Evaluator (DIIVINE) method is much more involved than the BLIINDS-II index, although it delivers more information regarding the quality of the distorted image. The division of QA tasks in DIIVINE makes it more useful for such important tasks as post-QA distortion reduction, but much less useful for real-time QA applications, as in a video network.

ENVISIONING THE FUTURE

Despite significant recent progress on the very old visual QA problems, there remains significant room for improvement. There is a gap in prediction performance between current performance and what we believe is possible (RR and NR IQA models that predict subject image quality as well as FR algorithms). There is a rather rich literature of NSS models, among which only a small proportion have been successfully exploited in the context of RR and NR QA [5], [18], [28].

A wide spectrum of novel functionalities could be added to RR/NR systems, making them more flexible and versatile in user-centric multimedia communication environments. One desirable feature is rate-scalability, wherein RR features are aligned (and possibly coded) to a continuous bit stream, and ordered according to importance. Such a bit stream could be truncated at any location, and the quality of the distorted images evaluated based on the truncated RR features. Quality prediction could be improved with increased length of the received bit stream. Ideally, such a rate-scalable method could

[TABLE 1] IQA ALGORITHM SCORES AGAINST HUMAN DMOS SCORES FROM LIVE IQA DATABASE [67]. SROCC OVER THE ENTIRE DATABASE FOR FR IQA INDICES PSNR, SS-SSIM, MS-SSIM, AND VIF. MEDIAN SROCC OVER 1,000 RANDOMIZED TRAIN-TEST SEQUENCES FOR NR IQA MODELS BLIINDS-II AND DIIVINE.

IQA ALGORITHM (RN MODELS IN BOLD)	JPEG 2000	JPEG	WHITE NOISE	GAUSSIAN BLUR	FAST FADING NOISE	ALL DATA
PSNR	0.90	0.83	0.99	0.78	0.89	0.87
SS-SSIM [10]	0.94	0.95	0.96	0.91	0.94	0.91
MS-SSIM [70]	0.97	0.96	0.98	0.95	0.94	0.95
VIF [6]	0.97	0.96	0.98	0.97	0.97	0.96
BLIINDS-II [56]	0.95	0.94	0.98	0.94	0.93	0.91
DIIVINE [59]	0.91	0.91	0.98	0.92	0.86	0.92

cover the full range of QA methods (NR, RR, and FR) within a unified framework. It would also be interesting to make the RR approach reverse-directional, where RR features are returned to the sender and compared with the reference features. This would be useful in a networking scenarios (e.g., broadcasting) where central quality control is at the sender side and the RR features from the receiver could help the sender make adaptive adjustments. Even further, one could design bidirectional RR systems, where the RR features could be sent either from the sender to the receiver or vice-versa.

AS PROGRESS CONTINUES, THE NEED FOR REFERENCE-FREE METHODS IS BECOMING MORE PRONOUNCED.

As progress continues, the need for reference-free methods is becoming more pronounced. This will be a very fast-growing area in the next five to ten years, driven by the needs of practical applications and by the many open problems that need to be solved. One of the most important problems to be solved, as we hinted at, is the RR/NR VQA problem, which will require the discovery of comprehensive video NSS models that are statistically regular, and that are sensitive to losses of naturalness induced by distortion. Another important problem is three-dimensional (3-D) stereoscopic image and video QA. Here also, there is a deficit of accurate and regular QA models, and good-performing algorithms (relative to two-dimensional algorithms applied to 3-D data) do not yet exist. We hope that this tutorial article can help attract and inspire more academic researchers and industrial practitioners to this fast-evolving field.

ACKNOWLEDGMENTS

The research of Alan C. Bovik was supported in part by Intel and Cisco Corporation under the VAWN Program and by the U.S. National Science Foundation under the IIS program. Zhou Wang was supported in part by the National Science and Engineering Research Council of Canada and by the Ontario Early Researcher Award program.

AUTHORS

Zhou Wang (zhouwang@ieee.org) is an associate professor in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include image processing and multimedia communications. He has more than 90 publications in these fields with over 7,000 citations. He was an associate editor of *IEEE Signal Processing Letters* (2006–2010). He is currently an associate editor of *Pattern Recognition* (2006–present) and *IEEE Transactions on Image Processing* (2009–present). He received the 2009 IEEE Signal Processing Best Paper Award, ICIP 2008 IBM Student Paper Award (as senior author), and 2009 Ontario Early Researcher Award.

Alan C. Bovik (bovik@ece.utexas.edu) is the Curry/Cullen Trust Chair Professor at The University of Texas at Austin. He is the director of LIVE in the Department of Electrical and Computer Engineering and the Institute for Neurosciences. His many awards include the 2011 IS&T Imaging Scientist of the Year Award and the 2009 IEEE Signal Processing Society Best Paper Award. He created the IEEE International Conference on Image Processing and cofounded *IEEE Transactions on Image Processing*. His books, articles on education, and award-winning online courseware and SIVA software attest to his dedication to engineering education.

REFERENCES

[1] Cisco Corporation. (Feb. 2011). Cisco visual networking index: Global mobile data traffic forecast update, 2010–2015. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf

[2] A. C. Bovik, "Meditations on video quality," *IEEE Multimedia Commun. E-Lett.*, vol. 4, no. 4, pp. 4–10, May 2009.

[3] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment.* San Rafael, CA: Morgan & Claypool, 2006.

[4] H. B. Barlow, "Possible principles underlying the transformation of sensory messages," in *Sensory Communications,* W. A. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, pp. 217–234.

[5] E. P. Simoncelli and B. Olshausen, "Natural image statistics and neural representation," *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, May 2001.

[6] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Processing*, vol. 15, pp. 430–444, Feb. 2006.

[7] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Trans. Image Processing*, vol. 12, pp. 1338–1351, Nov. 2003.

[8] H. R. Sheikh and A. C. Bovik, "An evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[9] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, and F. Battisti, "Color image database for evaluation of image quality metrics," in *Proc. Int. Workshop Multimedia Signal Processing*, Australia, Oct. 2008, pp. 403–408.

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[11] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 5, pp. 1185–1198, May 2011.

[12] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment systems based on human perception," *Proc. SPIE*, vol. 1913, pp. 15–26, 1993.

[13] B. Girod, "What's wrong with mean-squared error?" in *Visual Factors of Electronic Image Communications*. Cambridge, MA: MIT Press, 1993.

[14] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[15] Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet domain natural image statistic model," in *Proc. SPIE Conf. Human Vision Electronic Imaging*, Jan. 2005, vol. 5666, pp. 149–159.

[16] S. Mallat, *A Wavelet Tour of Signal Processing,* 2nd ed. San Diego, CA: Academic, 1999.

[17] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 1, pp. 55–73, Jan. 1990.

[18] D. J. Field, "What is the goal of sensory coding?" *Neural Comput.*, vol. 6, no. 4, pp. 559–601, 1994.

[19] W. S. Geisler and R. L. Diehl, "Bayesian natural selection and the evolution of perceptual systems," *Phil. Trans. R. Soc. Lond. B*, vol. 357, no. 1420, pp. 419–448, Apr. 2002.

[20] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, July 1989.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[22] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metric for in-service quality monitoring of any digital video system," *Proc. SPIE*, vol. 3845, pp. 266–277, Sept. 1999.

[23] I. P. Gunawan and M. Ghanbari, "Reduced-reference picture quality estimation by using local harmonic amplitude information," in *Proc. London Communications Symp.*, 2003, pp. 137–140.

[24] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. IEEE Int. Conf. Image Processing*, Barcelona, Spain, Sept. 2003, pp. 185–188.

[25] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Process. Image Commun.*, vol. 23, pp. 239–256, Apr. 2008.

[26] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Select. Topics Signal Process. (Special Issue on Visual Media Quality Assessment)*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[27] D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis. Neurosci.*, vol. 9, no. 2, pp. 181–198, 1992.

[28] D. L. Ruderman, "The statistics of natural images," *Network: Comput. Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1996.

[29] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of Gaussians and the statistics of natural images," *Adv. Neural Inform. Process. Syst.*, vol. 12, pp. 855–861, 2000.

[30] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Inform. Theory*, vol. 38, no. 2, pp. 587–607, Mar. 1992.

[31] J. Foley, "Human luminance pattern mechanisms: Masking experiments require a new model," *J. Opt. Soc. Amer.*, vol. 11, no. 6, pp. 1710–1719, 1994.

[32] A. C. Bovik, "What you see is what you learn," *IEEE Signal Processing Mag.*, vol. 27, no. 5, pp. 117–123, Sept. 2010.

[33] B. Girod, "The information theoretical significance of spatial and temporal masking in video signals," in *Proc. SPIE Conf. Human Vision, Visual Processing, Digital Display*, 1989, vol. 1077, pp. 178–187.

[34] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Processing*, Austin, TX, Nov. 1994, pp. 982–986.

[35] K. Seshadrinathan and A. C. Bovik, "Motion-tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[36] J. M. Kennedy, *A Psychology of Picture Perception*. San Francisco, CA: Jossey-Bass, 1974.

[37] I. van der Linde, U. Rajashekar, A. C. Bovik, and L. K. Cormack, "Visual memory for fixated regions of natural scenes dissociates attraction and recognition," *Perception*, vol. 38, no. 8, pp. 1152–1171, Aug. 2009.

[38] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE Int. Conf. Image Processing,* Rochester, NY, Sept. 2002, pp. 57–60.

[39] X. Zhu and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise," in *Proc. 1st Int. Workshop Quality of Multimedia Experience*, San Diego, CA, July 2009.

[40] Z. Wang, A. C. Bovik, and B. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Processing,* Vancouver, BC, Canada, Sept. 2000, pp. 981–984.

[41] S. Liu and A. C. Bovik, "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 12, pp. 1139–1149, 2002.

[42] Z. Wang, H. R. Sheikh and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. Image Processing.*, Rochester, NY, 2002, pp. 477–480.

[43] L. Meesters and J. Martens, "A single-ended blockiness measure for JPEG-coded images," *Signal Processing*, vol. 82, no. 3, pp. 369–387, 2002.

[44] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Applications to JPEG2000," *Signal Process. Image Commun.*, vol. 19, pp. 163–172, Feb. 2004.

[45] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Processing*, vol. 18, no. 4, pp. 717–728, Apr. 2009.

[46] R. Hassen, Z. Wang, and M. Salama, "No-reference image sharpness assessment based on local phase coherence measurement," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing,* Dallas, TX, Mar. 2010, pp. 2434–2437.

[47] H. R. Sheikh, A. C. Bovik, and L. K. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Processing*, vol. 14, no. 11, pp. 1918–1927, Nov. 2005.

[48] K. Tan and M. Ghanbari, "Blockiness detection for MPEG2-coded video," *IEEE Signal Processing Lett.*, vol. 7, no. 8, pp. 213–215, 2000.

[49] T. Vlachos, "Detection of blocking artifacts in compressed video," *Electron. Lett.*, vol. 36, no. 13, pp. 1106–1108, 2000.

[50] M. Ries, O. Nemethova, and M. Rupp, "Motion based reference-free quality estimation for H.264/AVC video streaming," in *Proc. Int. Symp. Wireless Pervasive Computing*, 2007, pp. 355–359.

[51] M. F. Sabir, R. W. Heath, and A. C. Bovik, "Joint source-channel distortion modeling for MPEG-4 video," *IEEE Trans. Image Processing*, vol. 18, no. 1, pp. 90–105, Jan. 2009.

[52] T. Brandao and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[53] S. Gabarda and G. Cristobal, "Blind image quality assessment through anisotropy," *J. Opt. Soc. Amer.*, vol. 24, no. 12, pp. B42–B51, 2007.

[54] M. A. Saad and A. C. Bovik, "A DCT statistics-based blind image quality index," *IEEE Signal Processing Lett.*, vol. 17, no. 6, pp. 583–586, June 2010.

[55] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, pp. 312–322, Sept. 2004.

[56] M. A. Saad, A. C. Bovik, and C. Charrier, "DCT statistics model-based blind image quality assessment," submitted for publication.

[57] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Special Topics Signal Processing*, vol. 3, pp. 193–201, Apr. 2009.

[58] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Processing Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[59] A. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," submitted for publication.

[60] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, no. 6, pp. 1680–1689, June 2006.

[61] J. H. Van Hateren and A. Van Der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc. R. Soc. Lond. B Biol. Sci.*, vol. 265, no. 1394, pp. 359–366, 1998.

[62] E. P. Simoncelli, "Capturing visual image properties with probabilistic models," in *The Essential Guide to Image Processing,* A. C. Bovik, Ed. San Diego, CA: Academic, 2009.

[63] J. H. van Hateren and D. L. Ruderman, "Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex," *Proc. R. Soc. Lond. B,* vol. 265, no. 1412, pp. 2315–2320, 1998.

[64] S. Roth and M. J. Black, "On the spatial statistics of optical flow," *Int. J. Comput. Vis.*, vol. 74, pp. 33–50, Aug. 2007.

[65] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEE Int. Conf Acoustics, Speech, Signal Processing,* Honolulu, HI, Apr. 2007.

[66] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[67] H. R Sheikh, Z. Wang, L. K. Cormack, and A. C. Bovik. (2005, Sept.). *LIVE Image Quality Database* [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm

[68] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "An evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[69] International Telecommunication Union. (2003). Methodology for the Subjective Assessment of the Quality for Television Pictures. ITU-R Rec. BT 500-11. [Online]. Available: http://www.dii.unisi.it/~menegaz/DoctoralSchool2004/papers/ITU-R_BT.500-11.pdf

[70] Z. Wang, E. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. Annu. Asilomar Conf. Signals, Systems, Computing, Pacific Grove*, CA, Nov. 2003, pp. 1398–1402.

[SP]

Christian Keimel, Martin Rothbucher, Hao Shen, and Klaus Diepold

Video Is a Cube

Multidimensional analysis and video quality metrics

Quality of experience (QoE) is becoming increasingly important in signal processing applications. In taking inspiration from chemometrics, we provide an introduction to the design of video quality metrics by using data analysis methods, which are different from traditional approaches. These methods do not necessitate a complete understanding of the human visual system (HVS). We use multidimensional data analysis, an extension of well-established data analysis techniques, allowing us to better exploit higher-dimensional data. In the case of video quality metrics, it enables us to exploit the temporal properties of video more properly; the complete three-dimensional structure of the video cube is taken into account in metrics' design. Starting with the well-known principal component analysis and an introduction to the notation of multiway arrays, we then present their multidimensional extensions, delivering better quality prediction results. Although we focus on video quality, the presented design principles can easily be adapted to other modalities and to even higher dimensional data sets as well.

A relatively new concept in signal processing, QoE aims to describe how video, audio, and multimodal stimuli are perceived by human observers. In the field of video quality assessment, it is often of interest for researchers how the overall experience is influenced by different video coding technologies, transmission errors, or general viewing conditions. The focus is no longer on measurable physical quantities, but rather on how the stimuli are subjectively experienced and whether they are perceived to be of acceptable quality from a subjective point of view.

QoE is in contrast to the well-established quality of service (QoS). There, we measure the signal fidelity, i.e., how much a signal is degraded during processing by noise or other disturbances. This is usually done by comparing the distorted with the original signal, which then gives us a measure of the signal's quality. To understand the reason why QoS is not sufficient for

Digital Object Identifier 10.1109/MSP.2011.942468
Date of publication: 1 November 2011

capturing the subjective perception of quality, let us take a quick look at the most popular metric in signal processing to measure the QoS: the mean squared error (MSE). It is known that the MSE does not correlate very well with the human perception of quality, as we just determine the difference between pixel values in both images. The example in Figure 1 illustrates this problem. Both images in Figure 1(a) have the same MSE with respect to the original image. Yet, we perceive the upper image distorted by coding artifacts to be of worse visual quality, than the lower image, where we just changed the contrast slightly. Further discussions of this problem can be found in [1].

A RELATIVELY NEW CONCEPT IN SIGNAL PROCESSING, QoE AIMS TO DESCRIBE HOW VIDEO, AUDIO, AND MULTIMODAL STIMULI ARE PERCEIVED BY HUMAN OBSERVERS.

HOW TO MEASURE QoE

How then can we measure QoE? The most direct way is to conduct tests with human observers, who judge the visual quality of video material and provide thus information about the subjectively perceived quality. However, we face a problem in real life: these tests are time-consuming and quite expensive. The reason for this is that only a limited number of subjects can take part in a test at the same time, but also because a multitude of different test cases have to be considered. Apart from these more logistical problems, subjective tests are usually not suitable if the video quality is required to be monitored in real time.

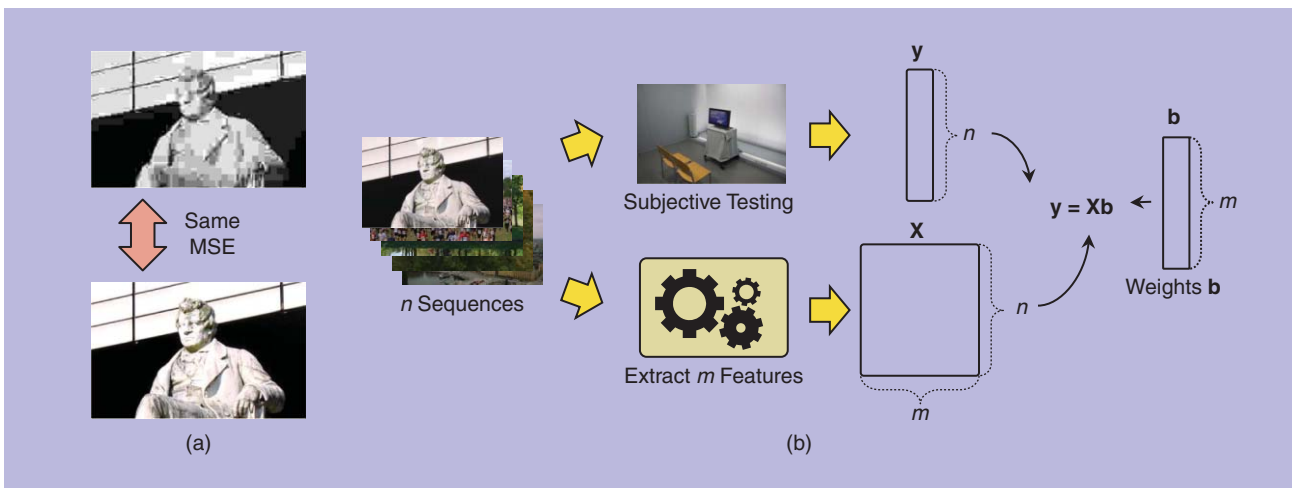
To overcome this difficulty, video quality metrics are designed and used. The aim is to approximate the human quality perception as satisfactory as possible with objectively measurable properties of the videos. Obviously, there is no single measurable quantity that by itself can represent the perceived QoE. Nevertheless, we can determine some aspects that are expected or shown to have a relation to the perception of quality and use these to design an appropriate video quality metric.

DESIGN OF VIDEO QUALITY METRICS—THE TRADITIONAL APPROACH

In the traditional approach, the quality metrics aim to implement the spatial and temporal characteristics of the HVS as well as possible in our metric. Many aspects of the HVS are not sufficiently understood and therefore a comprehensive model of the HVS is hardly possible to build. Nevertheless, at least parts of the HVS can be described sufficiently enough to utilize these properties in video quality metrics. In general, there are two different ways to exploit these properties according to Winkler [2]: either a psychophysical or an engineering approach.

The psychophysical approach relies primarily on a (partial) model of the HVS and tries to exploit known psychophysical effects, e.g., masking effects, adaption, and contrast sensitivity. One advantage of this approach is that we are not limited to a specific coding technology or application scenario, as we implement an artificial observer with properties of the HVS. In Daly's visual differences predictor [3], for example, the adaption of the HVS at different light levels is taken into account, followed by an orientation-dependent contrast sensitivity function, and finally, models of the HVS's different detection mechanism are applied. This full-reference predictor then delivers a measure for the perceptual difference in the same areas of two images. Further representatives of this approach are Lubin's visual discrimination model [4], the Sarnoff just noticeable difference (JND) [5], and Winkler's perceptual distortion metric [6].

In the engineering approach, the properties of the HVS are not implemented directly, but rather it determines features known to be correlated to the perceived visual quality. These features are then extracted and used in the video quality metric. As no in-depth understanding of the HVS is needed, this type of metric is commonly used in current research. In contrast to the psychophysical approach, however, we are limited to predefined coding technologies or application scenarios, as we do not



[FIG1] (a) Images with same MSE, but different visual quality, and how a model is built with data analysis; (b) subjective testing and feature extraction for each video sequence.

construct an artificial observer, but rather derive the features from artifacts introduced by the processing of the videos. One example for such a feature is blocking as seen in Figure 1. This feature is well known, as it is especially noticeable in highly compressed video. It is caused by block-based transforms such as the discrete cosine transform (DCT) or integer transform in the current video encoding standards MPEG-2 and H.264/AVC. With this feature, we exploit the knowledge that human perception is sensitive to edges, and therefore assume that artificial edges introduced by the encoding results in a degraded, perceived quality. Usually, more than one feature is extracted and these features are then combined into one value, by using assumptions about the HVS. Typical representatives of this approach are the widely used structural similarity (SSIM) index by Wang et al. [7], the video quality metric by Wolf and Pinson [8], and Watson's digital video quality metric [9]. Moreover we can distinguish between full-reference, reduced-reference, and no-reference metrics, where we have either the undistorted original, some meta information about the undistorted original, or only the distorted video available, respectively. We refer to [10] for further information on the HVS, and [11] and [12] for an overview of current video quality metrics.

In general, the exploitation of more properties of the HVS or their corresponding features in a metric allows us to model the perception of quality better. However, since the HVS is not understood completely, and consequently, no explicit model of the HVS describing all its aspects is available in the community, it is not obvious how the features should be combined. But do we really need to know a priori how to combine the features?

AN ALTERNATIVE METHOD: DATA ANALYSIS

Sometimes it is helpful to look at other disciplines. Video quality estimation is not the only application area in which we want to quantify something that is not directly accessible for measurement. Similar problems often occur in chemistry and related research areas. In food science, for example, researchers face a comparable problem: they want to quantify the taste of samples, but taste is not directly measurable. The classic example is about the determination of the perfect mixture for hot chocolate that tastes best. One can measure milk, sugar, or cocoa content, but there is not an a priori physical model that allows us to define the resulting taste. To solve this problem, a data-driven approach is applied, i.e., instead of making explicit assumptions of the overall system and relationship between the dependent variable, e.g., taste, and the influencing variables, e.g., milk, sugar, and cocoa, the input and output variable are analyzed. In this way we obtain models purely via the analysis of the data.

In chemistry, this is known as chemometrics and has been applied successfully to many problems in this field for the last three decades. It provides a powerful tool to tackle the analysis and prediction of systems that are understood only to a limited degree. So far, this method is not well known in the context of video quality assessment or even multimedia quality assessment in general. A good introduction into chemometrics can be found in [13].

By applying this multivariate data analysis to video quality, we now consider the HVS as a black box and therefore do not assume a complete understanding of it. The input corresponds to features we can measure and the output of the box to the perceived visual quality obtained in subjective tests. First, we extract m features from an image or video frame I, resulting in a $1 \times m$ row vector $\mathbf{x}$. While this is similar to the engineering approach described in the previous section, an important difference is that we do not make any assumption about the relationship between the features themselves, but also not about how they are combined into a quality value.

In general, we should not limit the number of selected features unnecessarily. Or to quote Martens and Martens [13], "Beware of wishful thinking!" As we do not have a complete understanding of the underlying system, it can be fatal if we exclude some features before conducting any analysis, because we consider them to be irrelevant. On the other hand, data that can be objectively extracted, like the features in our case, is usually cheap or in any case less expensive to generate than subjective data gained in tests. If some features are irrelevant to the quality, we will find out during the analysis. Of course it is only sensible to select features that have some verified or at least some suspected relation to the human perception of visual quality. For example, we could measure the room temperature, but it is highly unlikely that room temperature has any influence in our case.

For n different video sequences, we extract a corresponding feature vector $\mathbf{x}$ for each sequence and thus get an $n \times m$ matrix $\mathbf{X}$, where each row describes a different sequence or sample and each column describes a different feature as shown in Figure 1. We generate a subjective quality value for each of the n sequences by subjective testing and get an $n \times 1$ column vector $\mathbf{y}$ that will represent our ground truth. Based on this data set, a model can be generated to explain the subjectively perceived quality with objectively measurable features. Our aim is now to find an $m \times 1$ column vector $\mathbf{b}$ that relates the features in $\mathbf{X}$ to our ground truth in $\mathbf{y}$ or provides the weights for each feature to get the corresponding visual quality. This process is called calibration or training of the model, and the used sequences are the training set. We can use $\mathbf{b}$ to also predict the quality of new, previously unknown sequences. The benefit of using this approach is that we are able to combine totally different features into one metric without knowing their proportional contribution to the overall perception of quality beforehand.

CLASSIC AND WELL KNOWN: LINEAR REGRESSION

One classic approach to estimate the weight vector $\mathbf{b}$ is via a simple multiple linear regression model, i.e.,

$$\mathbf{y} = \mathbf{X}\hat{b} + \epsilon, \tag{1}$$

where ϵ is the error term. Without loss of generality, the data matrix $\mathbf{X}$ can be assumed to be centered, namely with zero means, and consequently the video quality values $\mathbf{y}$ are also centered.

Using a least squares estimation, we are given an estimation of $\mathbf{b}$ as

$$\hat{\mathbf{b}} = (\mathbf{X}^\top \mathbf{X})^+ \mathbf{X}^\top \mathbf{y}, \tag{2}$$

where Z^+ denotes the More-Penrose pseudo-inverse of matrix Z. We use the pseudo-inverse, as we can not assume that columns of $\mathbf{X}$ representing the different features are linearly independent and therefore $\mathbf{X}^\top\mathbf{X}$ can be rank deficient. For an unknown video sequence V_U and the corresponding feature vector $\mathbf{x}_u$, we are then able to predict its visual quality $\hat{y}_u$ with

$$\hat{y}_u = \mathbf{x}_u \hat{\mathbf{b}}. \tag{3}$$

Yet, this simple approach has a drawback: we assume implicitly in the estimation process of the weights that all features are equally important. Clearly, this will not always be the case, as some features may have a larger variance than others.

AN IMPROVEMENT: PRINCIPAL COMPONENT REGRESSION

We can address the aforementioned issue by selecting the weights in the model, so that they take into account the influence of the individual features on the variance in the feature matrix $\mathbf{X}$. We are therefore looking for so-called latent variables, that are not directly represented by the measured features themselves, but rather by a hidden combination of them. In other words, we aim to reduce the dimensionality of our original feature space into a more compact representation, more fitting for our latent variables. One well known method is the principal component analysis (PCA), which extracts the latent variables as the principal components (PCs). The variance of the PCs is expected to preserve the variance of the original data. We then perform a regression on some of these PCs leading to the PC regression (PCR). As PCA is a well-known method, we just briefly recap some basics.

Let $\mathbf{X}$ be a (centered) data matrix, we define $r = \min\{n, m\}$ and using a singular value decomposition (SVD) we get the following factorization:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{P}^\top, \tag{4}$$

where $\mathbf{U}$ is an $n \times r$ matrix with r orthonormal columns, $\mathbf{P}$ is an $m \times r$ matrix with r orthonormal columns, and $\mathbf{D}$ is a $r \times r$ diagonal matrix. $\mathbf{P}$ is called the loadings matrix, and its columns $\mathbf{p}_1, \ldots, \mathbf{p}_r$ are called loadings. They represent the eigenvectors of $\mathbf{X}^\top\mathbf{X}$. Furthermore we define the scores matrix

$$\mathbf{T} = \mathbf{U}\mathbf{D} = \mathbf{X}\mathbf{P}. \tag{5}$$

The basic idea behind PCR is to approximate $\mathbf{X}$ by only using the first g columns of $\mathbf{T}$ and $\mathbf{P}$, representing the g largest eigenvalues of $\mathbf{X}^\top\mathbf{X}$ and also the first g PCs. We hereby assume that the combination of the largest g eigenvalues describe the variance in our data matrix $\mathbf{X}$ sufficiently and that we can therefore discard the smaller eigenvalues. If g is smaller than r, the model can be built with a reduced rank. Usually we aim to explain at least 80–90% of the variance in $\mathbf{X}$. But other selection criteria are also possible.

Our regression model with the first g PCs can thus be written as

$$\mathbf{y} = \mathbf{T}_g \mathbf{c}, \tag{6}$$

where $\mathbf{T}_g$ represents a matrix with the first g columns of $\mathbf{T}$ and $\mathbf{c}$ the (unknown) weight vector. Once again, we perform a multiple linear regression. We estimate $\mathbf{c}$ with the least squares method as

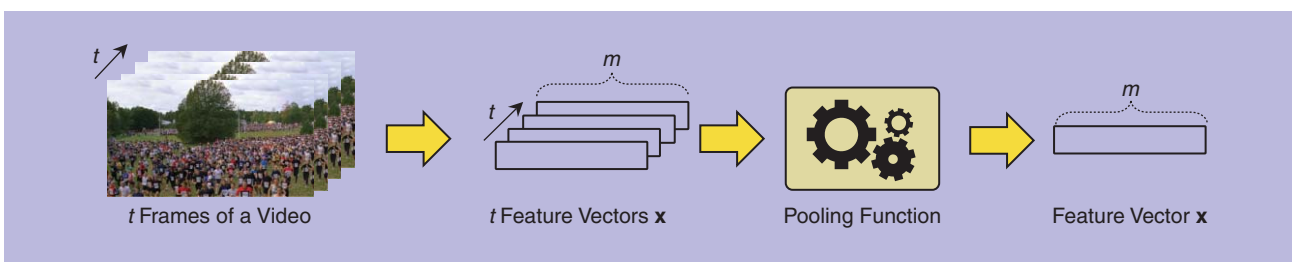
$$\hat{\mathbf{c}} = (\mathbf{T}_g^\top \mathbf{T}_g)^{-1} \mathbf{T}_g^\top \mathbf{y}. \tag{7}$$

In the end, we are interested in the weights, so that we can directly calculate the visual quality. Therefore we determine the estimated weight vector $\hat{\mathbf{b}}$ as

$$\hat{\mathbf{b}} = \mathbf{P}_g \hat{\mathbf{c}}, \tag{8}$$

with $\mathbf{P}_g$ representing the matrix with the first g columns of $\mathbf{P}$. We can predict the visual quality for an unknown video sequence V_U and the corresponding feature vector $\mathbf{x}_u$ with (3).

PCA was first used in the design of video quality metrics by Miyahara in [14]. We refer to [15] for further information on PCA and PCR. A more sophisticated method often used in chemometrics is the partial least squares regression (PLSR). This method also takes the variance in the subjective quality vector $\mathbf{y}$ into account as well as the variance in the feature matrix $\mathbf{X}$. PLSR has been used in the design of video quality metrics in e.g., [16]. Further information on PLSR itself can be found in [13] and [17].



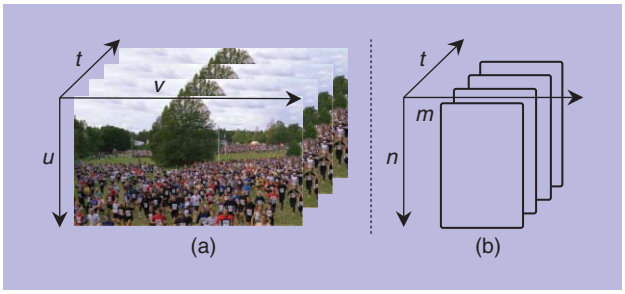
[FIG2] **Temporal pooling.**

VIDEO IS A CUBE

The temporal dimension is the main difference between still images and video. In the previous section we assumed that we extract the feature vector only from one image or one video frame, which is a two-dimensional (2-D) matrix. In other words, video was considered to be just a simple extension of still images. This is not a unique omission only in this article so far, but the temporal dimension is quite often neglected in many contributions in the field of video quality metrics. The additional dimension is usually managed by temporal pooling. Either the features themselves are temporally pooled into one feature value for the whole video sequence or the metric is applied to each frame of the video separately and then the metric's values are pooled temporally over all frames to gain one value, as illustrated in Figure 2.

Pooling is mostly done by averaging, but also other simple statistical functions are employed such as standard deviation, 10/90% percentiles, and median or minimum/maximum. Even if a metric considers not only the current frame, but also preceding or succeeding frames, e.g., with a three-dimensional (3-D) filter [18] or spatiotemporal tubes [19], the overall pooling is still done with one of the above functions. But this arbitrary pooling, especially averaging, obscures the influence of temporal distortions on the human perception of quality, as intrinsic dependencies and structures in the temporal dimension are disregarded. The importance of video features' temporal properties in the design of video quality metrics was recently shown in [20].

Omitting the temporal pooling step and introducing the additional temporal dimension directly in the design of the video quality metrics can improve the prediction performance. We propose therefore to consider video in its natural 3-D structure as a video cube. Extending the data analysis approach, we add an additional dimension to our data set and thus arrive at multidimensional data analysis, an extension of the two dimensional data analysis. In doing so, we gain a better understanding of the video's properties and will thus be able to interpret the extracted features better. We no longer employ an a priori temporal pooling step but use the whole video cube to generate the prediction model for the visual quality, and thus consider the temporal dimension of video more appropriately.



[FIG3] (a) Video cube and (b) feature cube.

TENSOR NOTATION

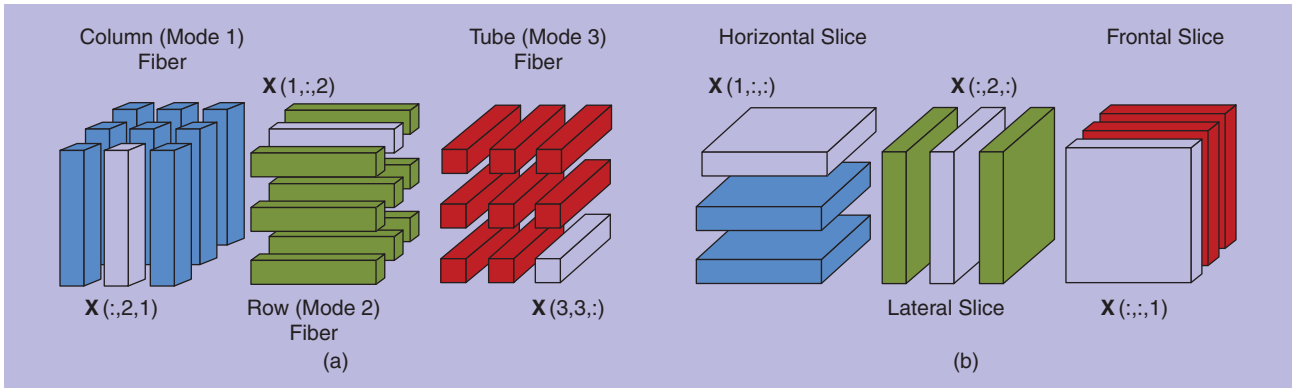
Before moving on to the multidimensional data analysis, we shortly introduce the notation for handling multiway arrays or tensors.

In general, our video cube can be presented as a three-way $u \times v \times t$ array $\mathbf{V}(:,:,:)$, where the u and v are the frame size, and t is the number of frames. Similarly, we can extend the 2-D feature matrix $\mathbf{X}$ into the temporal dimension as a $n \times m \times t$ three-way array or feature cube. Both are shown in Figure 3. In this work, we denote $\mathbf{X}(i,j,k)$ as the (i,j,k)th entry of $\mathbf{X}$, $\mathbf{X}(i,j,:)$ as the vector with a fixed pair of (i,j) of $\mathbf{X}$, referred to as tensor fiber, and $\mathbf{X}(i,:,:)$ the matrix of $\mathbf{X}$ with a fixed index i, referred to as tensor slice. The different fibers and slices are shown in Figure 4. For more information about tensors and multiway arrays, see [21], and for multiway data analysis, refer to [22] and [23].

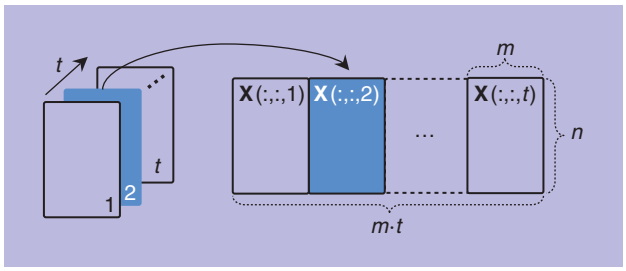
UNFOLDING TIME

The easiest way to apply conventional data analysis methods for analyzing tensor data is to represent tensors as matrices. It transforms the elements of a tensor or multiway array into entries of a matrix. Such a process is known as unfolding, matricization, or flattening.

In our setting, we are interested in the temporal dimension and therefore perform the Mode-1 unfolding of our three-way array $\mathbf{X}(i,j,k)$. Thus, we obtain a new $n \times (m \cdot t)$ matrix $\mathbf{X}_{\text{unfold}}$, whose columns are arranged Mode-1 fibers of $\mathbf{X}(i,j,k)$. For simplicity, we assume that the temporal order is maintained in $\mathbf{X}_{\text{unfold}}$. The structure of this new matrix is shown in Figure 5.



[FIG4] Tensor notation: (a) fiber and (b) slice.



[FIG5] **Unfolding of the feature cube.**

We then perform a PCR on this matrix as described previously and obtain a model of the visual quality. Finally, we can predict the visual quality of an unknown video sequence V_U with its feature vector $\mathbf{x}_u$ by using (3).

Note, that $\mathbf{x}_u$ is now of the dimension $1 \times (m \cdot t)$. One disadvantage during the model building step with PCR is that the SVD must be performed on a rather large matrix. Depending on the frames in the video sequence, the time needed for model building can increase by a factor of 10^3 or higher. But more importantly, we still lose some information about the variance by unfolding and thus destroying the temporal structure.

2-D PRINCIPAL COMPONENT REGRESSION

Instead of unfolding, we can include the temporal dimension directly in the data analysis via performing a multidimensional data analysis. We use the 2-D extension of the PCA, (2-D-PCA), recently proposed by Yang et al. [24], in combination with a least squares regression as 2-D-PCR. For a video sequence with t frames, we can carve the $n \times m \times t$ feature cube into t slices, where each slice represents one frame. Without loss of generality, we can compute the covariance or scatter matrix as

$$\mathbf{X}_{Sct} = \frac{1}{t}\sum_{i=1}^{t}\mathbf{X}(:,:,i)^{\top}\mathbf{X}(:,:,i), \tag{9}$$

where, by abusing the notation, $\mathbf{X}(:,:,i)$ denotes the centered data matrix. It describes therefore the average covariance over the temporal dimension t. Then we perform the SVD performed on $\mathbf{X}_{Sct}$ to extract the PCs, similar to the previously described one dimensional PCR in (4).

Instead of a scores matrix $\mathbf{T}$, we now have a three-way $n \times g \times t$ scores array $\mathbf{T}(:,:,:)$, with each slice defined as

$$\mathbf{T}(:,:,i) = \mathbf{X}(:,:,i)\mathbf{P}. \tag{10}$$

Similar to (7), we then estimate a $g \times 1 \times t$ prediction weight for each slice with the first g principal components as

$$\hat{\mathbf{C}}(:,:,i) = \left(\mathbf{T}_g(:,:,i)^{\top}\mathbf{T}_g(:,:,i)\right)^{+}\mathbf{T}_g(:,:,i)^{\top}\mathbf{y}(i), \tag{11}$$

before expressing the weights in our original feature space with a $m \times 1 \times t$ three-way array

$$\hat{\mathbf{B}}(:,:,i) = \mathbf{P}_g\hat{\mathbf{C}}(:,:,\mathbf{i}), \tag{12}$$

comparable to (8) for the one-dimensional PCR. Note, that the weights are now represented by a (rotated) matrix.

A quality prediction for the ith slice can then be performed in the same manor as in (3), i.e.,

$$\hat{\mathbf{y}}_u(i) = \mathbf{X}_u(:,:,i)\hat{\mathbf{B}}(:,:,i), \tag{13}$$

where $\mathbf{X}_u$ represents a $1 \times m \times t$ feature matrix for one sequence and $\hat{\mathbf{y}}_u(i)$ the $1 \times t$ predicted quality vector. We can now use this quality prediction individually for each slice or generate one quality value for the whole video sequence by pooling. So far, 2-D-PCR has been used for video quality metrics in [25].

CROSS VALIDATION

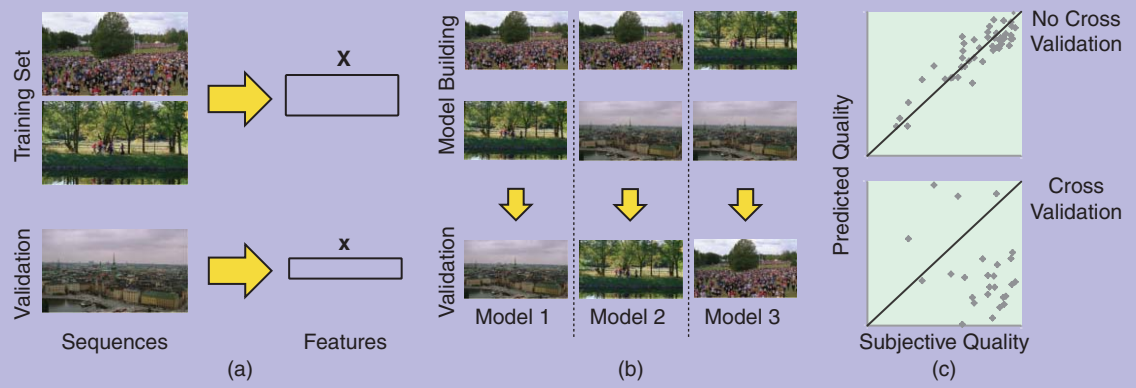
In general, data analysis methods require a training phase or a training set. One important aspect is to employ a separate data set for the validation of the designed metric. Using the same data set for training and validation will usually give us misleading results. Not surprisingly, our metric performs excellently with its training data. For unknown video sequences, on the other hand, the prediction quality could be very bad. But as mentioned previously, the data in video quality metrics is usually expensive to generate as we have to conduct subjective tests. Hence we can not really afford to use only a subset of all available data for the model building, as the more training data we have, the better the prediction abilities of our metric will be.

This problem can be partially avoided by performing a cross validation, e.g., leave-one-out. This allows us to use all available data for training, but also to use the same data set for the validation of the metric. Assuming we have i video sequences with different content, then we use $i - 1$ video sequences for training and the left out sequence for validation. All in all, we eventually get i models. This is illustrated in Figure 6. The general model can then be obtained by different methods, e.g., averaging the weights over all models or selecting the model with the best prediction performance. For more information on cross validation in general, we refer to [13] and [26].

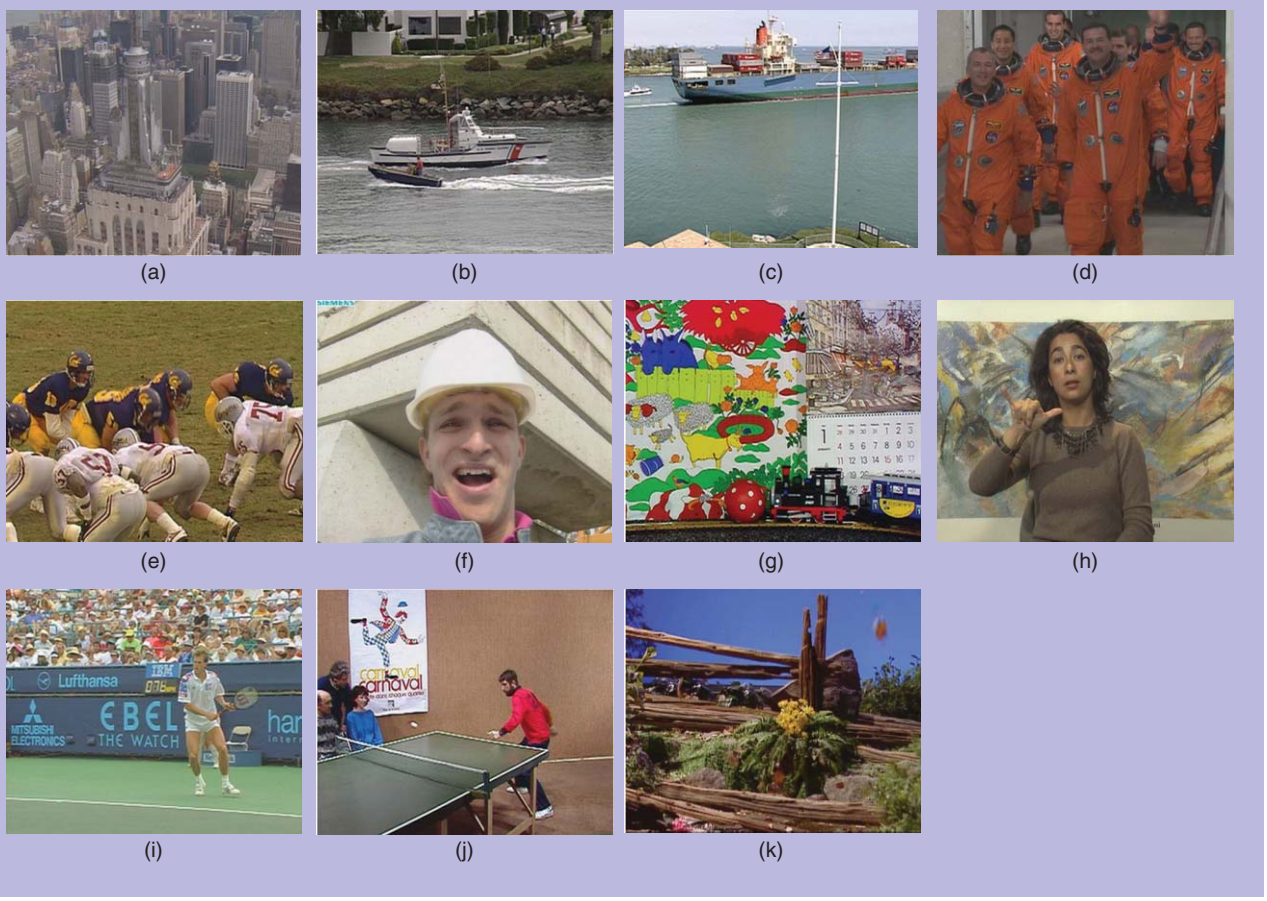
USING DATA ANALYSIS: AN EXAMPLE METRIC

But do more dimensions really help us in designing better video quality metrics? To compare the approaches to data analysis we presented in this work, we therefore design as an example a simple metric for estimating the visual quality of coded video with each method in this section.

The bit stream of an encoded video provides easily accessible objective data. Even though we do not know a priori which of the bit stream's properties are more or less important, we can safely assume that they are related in some way to the perceived visual quality. How they are related will be determined by data analysis. In this example, we use videos encoded with the popular H.264/AVC standard, currently used in many applications from high-definition TV (HDTV) to Internet-based IPTV. For each frame, we extract 16 different features describing the



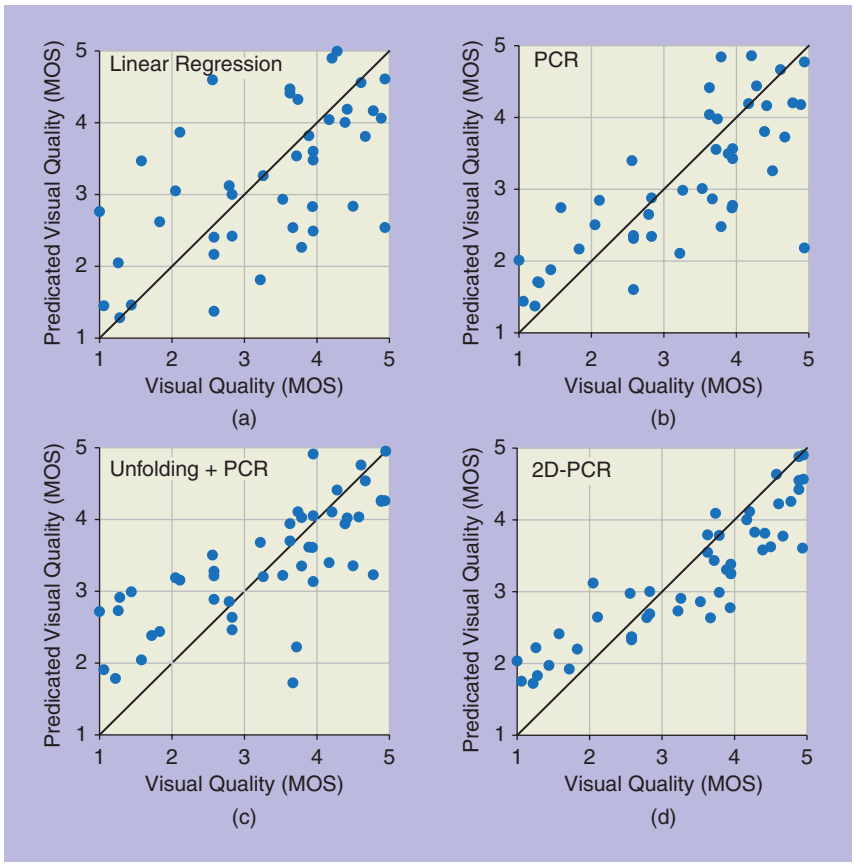
[FIG6] Cross validation: (a) splitting up the data set in training and validation set, (b) building different models with each combination, and (c) quality prediction of a metric.



[FIG7] Test videos: (a) City; (b) Coast Guard; (c) Container; (d) Crew; (e) Football; (f) Foreman; (g) Mobile; (h) Silent; (i) Stephan; (j) Table Tennis; and (k) Tempete.

partitioning into different block sizes and types, the properties of the motion vectors and lastly the quantization, similar to the metric proposed in [27]. Each frame is thus represented as 1×16 feature vector $\mathbf{x}$. Note, that no further preprocessing of the bit stream features was done. Alternatively, one can also extract features independent of the used coding technology, e.g., blocking or blurring, as described in [16].

Certainly, we also need subjective quality values for these encoded videos as ground truth to perform the data analysis. Different methodologies as well as the requirements on the test set-up and equipment for obtaining this data are described in international standards, e.g., ITU-R BT.500 or ITU-T P.910. Another possibility is to use existing, publicly available data sets, containing both the encoded videos and the visual quality



[FIG8] **Comparison of the presented methods (a)–(d): subjective quality versus predicated quality on a mean opinion score (MOS) scale from 1 to 5, worst to best quality.**

[TABLE 1] PERFORMANCE MEASUREMENTS: PEARSON CORRELATION, SPEARMAN RANK ORDER CORRELATION, AND RMSE. ADDITIONALLY, THE RATIO OF HOW MANY QUALITY PREDICTIONS ARE OUTSIDE OF THE GIVEN SCALE.

	PEARSON CORRELATION	SPEARMAN CORRELATION	RMSE	OUTSIDE SCALE
LINEAR REGRESSION	0.72	0.72	1.04	15%
PCR	0.80	0.82	0.81	13%
UNFOLDING + PCR	0.75	0.83	0.82	10%
2-D-PCR	0.89	0.94	0.59	6%

values. One advantage of using such data sets is that different metrics can be compared more easily.

For this example, we will use a data set provided by IT-IST [28]. It consists of 11 videos in common intermediate format (CIF) resolution (352×288) and a frame rate of 30 frames/s as shown in Figure 7. They cover a wide range of different content types, at bitrates from 64 kb/s to 2,000 kb/s, providing a wide visual quality range with in total $n = 52$ data points, leading to a 52×1 quality vector $\mathbf{y}$. According to [28], the test was conducted using to the DCR double stimulus method described in ITU-T P.910. For each data point, the test subjects were shown the undistorted original video, followed by the distorted encoded video and then asked to assess the impairment of the coded video with respect to the original on a discrete five-point mean opinion score (MOS) scale from one (very annoying) to five (imperceptible). For more information on H.264/AVC in general, we refer to [29], and for the H.264/AVC feature extraction to [27]. A comprehensive list of publicly available data sets is provided in [30].

MORE DIMENSIONS ARE REALLY BETTER

Finally, we compare the four video quality metrics, each designed with one of the presented methods. By using a cross-validation approach, we design 11 different models for each method. Each model is trained using ten video sequences and the left out sequence is then used for validation of the model built with the training set. Hence, we can measure the prediction performance of the models for unknown video sequences.

The performance of the different models is compared by calculating the Pearson correlation and the Spearman rank order correlation between the subjective visual quality and the quality predictions. The Pearson correlation gives an indication about the prediction accuracy of the model and the Spearman rank order correlation gives an indication how much the ranking between the sequences changes between the predicted and subjective quality. Additionally, we determine the root mean squared error (RMSE) between prediction and ground truth, but also the percentage of predictions that fall outside the used quality scale from one to five.

By comparing the results in Figure 8 and Table 1, we can see that a better inclusion of the temporal dimension in the model building helps to improve the prediction quality. Note that this improvement was achieved very easily, as we did nothing else, but just changing the data analysis method. In each step, we exploit the variation in our data better. First just within our temporally pooled features with the step from multiple linear regression to PCR, then by the step in the third dimension with unfolding and 2-D-PCR.

SUMMARY

In this work, we provide an introduction into the world of data analysis and especially the benefits of multidimensional data analysis in the design of video quality metrics. We have seen in our example, that even with a very basic metric, we can increase the performance of predicting the QoE significantly, if we use

multidimensional data analysis. Although the scope of this introduction covered only the quality of video, the proposed methods can obviously be extended to more dimensions and/or other areas of application. It is interesting to note that the dimensions need not be necessarily spatial or temporal, but also may represent different modalities or perhaps even a further segmentation of the existing feature spaces.

AUTHORS

Christian Keimel (christian.keimel@tum.de) received the B.Sc. and Dipl.-Ing. degrees in electrical engineering and information technology from the Technische Universität München (TUM) in 2005 and 2007, respectively. He is currently pursuing a Ph.D. degree at the Institute for Data Processing at TUM. His research interests include video quality assessment and the application of multidimensional data analysis methods in the context of QoE. He is also a deputy work group leader in the European network on QoE in multimedia systems and services (QUALINET), where he is focusing on future application areas and use cases for QoE.

Martin Rothbucher (martin.rothbucher@tum.de) received the Dipl.-Berufspäd. degree in electrical engineering and information technology from TUM in 2008. He is currently pursuing a Ph.D. degree at the Institute for Data Processing at TUM. His research interests include audio signal processing and applications of multidimensional data analysis methods, particularly in the context of head-related transfer functions. He is also involved in the collaborative research center SFB453 "High-Fidelity Telepresence and Teleaction" of the German research foundation (DFG) with a focus on acoustic telepresence.

Hao Shen (hao.shen@tum.de) received the bachelor's degree in mechanical engineering and applied mathematics from Xi'an Jiaotong University, China, in 2000, master's degrees in computer studies and computer science from the University of Wollongong, Australia, in 2002 and 2004, respectively, and the Ph.D. degree from the Australian National University, in 2008. Currently, he is a postdoctoral researcher at the Institute for Data Processing at TUM, Germany. His research interests focus on autonomous learning, geometric optimization, blind source separation, and related topics in signal processing.

Klaus Diepold (kldi@tum.de) received a Dipl.-Ing. and a Dr.-Ing. degree both in electrical engineering from TUM in 1987 and 1992, respectively. He spent more than ten years in the media industry being actively involved in MPEG standardization. In 2002, he joined TUM as professor with the Department of Electrical Engineering and Information Technology. His main research interests are in audiovisual signal processing, data analysis, and machine learning.

REFERENCES

[1] Z. Wang and A. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Mag.*, vol. 26, no. 1, pp. 98–117, Jan. 2009.

[2] S. Winkler, "Perceptual video quality metrics—A review," in *Digital Video Image Quality and Perceptual Coding,* H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC, 2006, pp. 155–179.

[3] S. J. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision,* A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 179–206.

[4] J. Lubin, "A visual discrimination model for imaging system design and evaluation," *in Vision Models for Target Detection and Recognition,* A. R. Menendez and E. Peli, Eds. Singapore: World Scientific, 1995, pp. 245–283.

[5] J. Lubin and D. Fibush, *Sarnoff JND Vision Model, T1A1.5 Working Group, ANSI T1 Standards Committee Standard*, 1997.

[6] S. Winkler, *Digital Video Quality—Vision Models and Metrics*. Hoboken, NJ: Wiley, 2005.

[7] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[8] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metric for in-service quality monitoring of any digital video system," in *Proc. Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Multimedia Systems and Applications II*, Nov. 1999, vol. 3845, pp. 266–277.

[9] A. B. Watson, "Toward a perceptual video-quality metric," in *Proc. Society of Photo-Optical Instrumentation Engineers (SPIE) Conf. Human Vision and Electronic Imaging III*, Jan. 1998, vol. 3299, pp. 139–147.

[10] B. A. Wandell, *Foundations of Vision.* Sunderland, MA*:* Sinauer Associates, 1996.

[11] H. R. Wu and K. R. Rao, Eds., *Digital Video Image Quality and Perceptual Coding*. Boca Raton, FL: CRC, 2006.

[12] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment (Synthesis Lectures on Image, Video, and Multimedia Processing)*. San Rafael, CA: Morgan & Claypool Publishers, 2006.

[13] H. Martens and M. Martens, *Multivariate Analysis of Quality*. New York: Wiley, 2001.

[14] M. Miyahara, "Quality assessments for visual service," *IEEE Commun. Mag.*, vol. 26, no. 10, pp. 51–60, 81, Oct. 1988.

[15] I. Jolliffe, *Principal Component Analysis*. Berlin: Springer-Verlag, 2002.

[16] T. Oelbaum, C. Keimel, and K. Diepold, "Rule-based no-reference video quality evaluation using additionally coded videos," *IEEE J. Select. Topics Signal Processing*, vol. 3, no. 2, pp. 294–303, Apr. 2009.

[17] F. Westad, K. Diepold, and H. Martens, "QR-PLSR: Reduced-rank regression for high-speed hardware implementation," *J. Chemometrics*, vol. 10, no. 5–6, pp. 439–451, 1996.

[18] K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Processing*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[19] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Select. Topics Signal Processing,* vol. 3, no. 2, pp. 253–265, Apr. 2009.

[20] C. Keimel, T. Oelbaum, and K. Diepold, "Improving the prediction accuracy of video qualtiy metrics," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP* 2010), Mar. 2010, pp. 2442–2445.

[21] A. Cichocki, R. Zdunek, A. H. Phan, and S.-I. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Hoboken, NJ: Wiley, 2009.

[22] A. Smilde, R. Bro, and P. Geladi, *Multi-Way Analysis: Applications in the Chemical Sciences*. Hoboken, NJ: Wiley, 2004.

[23] P. M. Kroonenberg, *Applied Multiway Data Analysis*. Hoboken, NJ: Wiley, 2008.

[24] J. Yang, D. Zhang, A. Frangi, and J. Yang, "Two-dimensional PCA: A new approach to appearance-based face representation and recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 1, pp. 131–137, Jan. 2004.

[25] C. Keimel, M. Rothbucher, and K. Diepold, "Extending video quality metrics to the temporal dimension with 2D-PCR," *in Image Quality and System Performance VIII*, vol. 7867, S. P. Farnand and F. Gaykema, Eds. Bellingham, WA: SPIE, Jan. 2011.

[26] E. Anderssen, K. Dyrstad, F. Westad, and H. Martens, "Reducing over-optimism in variable selection by cross-model validation," *Chemometrics Intell. Lab. Syst.*, vol. 84, no. 1–2, pp. 69–74, 2006.

[27] C. Keimel, J. Habigt, M. Klimpke, and K. Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *Proc. IEEE Int. Workshop Quality of Multimedia Experience (QoMEX'11)*, Sept. 2011, pp. 49–54.

[28] T. Brandão and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, Nov. 2010.

[29] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, July 2003.

[30] S. Winkler. (2011, July). Image and video quality resources. [Online]. Available: http://stefan.winkler.net/resources.html

[SP]

[Ulrich Engelke, Hagen Kaprykowsky, Hans-Jürgen Zepernick, and Patrick Ndjiki-Nya]

Visual Attention in Quality Assessment

[Theory, advances, and challenges]

Perceptual quality metrics are widely deployed in image and video processing systems. These metrics aim to emulate the integral mechanisms of the human visual system (HVS) to correlate well with visual perception of quality. One integral property of the HVS is, however, often neglected: visual attention (VA) [1]. The essential mechanisms associated with VA consist mainly of higher cognitive processing, deployed to reduce the complexity of scene analysis. For this purpose, a subset of the visual information is selected by shifting the focus of attention across the visual scene to the most relevant objects. By neglecting VA, perceptual quality models inherently assume that all objects draw the attention of the viewer to the same degree. This applies to both the natural scene content as well as possibly induced distortions. However, suprathreshold distortions can be a strong attractor of VA and as a result, have a severe impact on the perceived quality. Identifying the perceptual influence of distortions relative to the natural content can thus be expected to enhance the prediction performance of perceptual quality metrics. The potential benefit of integrating VA information into image and video quality models has recently been recognized by a number of research groups [2]–[20]. The conclusions drawn from these works are somewhat controversial and give rise to many open questions. The goals of this article are therefore to shed some light onto this immature research field and to provide guidance for further advances. Toward these goals, we first discuss VA concepts that are relevant in the context of quality perception. We then review recent advances in research on integrating VA into quality assessment and

Digital Object Identifier 10.1109/MSP.2011.942473
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE

highlight the main findings. Finally, we discuss major challenges and suggest potential solutions and future directions.

VISUAL ATTENTION

The human eye faces an abundant amount of visual information at any instant in time. Several mechanisms in early vision and higher cognitive layers are therefore deployed to reduce the complexity of scene analysis.

RETINAL SAMPLING AND EYE MOVEMENTS

Nonuniform sampling is deployed on the retina with a high sampling density in the fovea and rapidly diminishing density with increasing eccentricity. Hence, high-accuracy processing is limited to the central focus point, the fovea, and the peripheral visual field is perceived with lower accuracy.

A visual scene is gradually inspected by shifting the focus point using rapid, saccadic eye movements to fixate on the most relevant information in any context. Visual perception is active only during fixations and is largely suppressed during saccades [21]. Even though visual scene sampling is dominated by fixations, the scene is perceived as a continuous visual world. Fixations on moving objects are enabled through smooth pursuit eye movements, during which high acuity processing is performed for object speeds of up to approximately two degrees of visual angle per second [22].

VISUAL ATTENTION MECHANISMS

VA is thought to have evolved as a result of the limited overall resources available in the HVS [23]. Visual stimuli are therefore constantly competing for these resources and the most relevant stimuli in a given context are favored over the less relevant ones. The quality and acuity of the attended stimulus is enhanced through increased gain and contrast sensitivity, accompanied by a widespread baseline-activity reduction and noise suppression in the remaining visual field [24]. The decision which stimuli are favored is influenced by a number of different mechanisms.

OVERT VERSUS COVERT ATTENTION

Eye movements do not necessarily reflect exactly what human observers are attending to [25]. Overt VA relates to the act of directing the eyes to a stimulus whereas covert VA is related to a mental shift of attention. Covert attention precedes eye movements [26] and during fixation, it can be deployed to multiple locations simultaneously. Hence, covert VA allows us to efficiently monitor the visual scene and guide our eye movements. It is even possible to pursue one target while attending another target with only little effect on the pursuit [27]. Overt and covert VA are strongly interlinked and thus, eye tracking experiments are widely used to measure overt VA of human observers to gain insights into the attentive behavior.

SPATIAL, FEATURE-BASED, AND OBJECT-BASED ATTENTION

VA is strongly influenced by three cues that are deployed simultaneously in a mutually optimal way; spatial location, low-level features, and objects [23]. Overt spatial attention is accompanied by eye movements whereas covert spatial attention can be deployed in the peripheral visual field and is thus not directly observable. Feature-based attention is largely independent of location and is affected by low-level features that are visually salient, including color, motion, orientation, and size [25]. It is active simultaneously throughout the visual field and is thus instrumental in improving detection performance of relevant stimuli. Object-based attention is guided by higher-level features, such as object structures as well as semantic information and contextual effects. Context plays a particularly important role in the decision process as to which object is considered more relevant than others [26].

BOTTOM-UP AND TOP-DOWN MECHANISMS

VA is guided by two main mechanisms: bottom-up and top-down. The former is reflexive, signal driven, and independent of a particular task. Bottom-up attention is fast, short lasting (transient), and performed in a preattentive manner across the visual field. It is driven involuntarily as a response to certain low-level features that are experienced as visually salient and distinct from the background. Motion, and in particular sudden temporal changes, are known to be dominant features in dynamic visual scenes [28], [29]. Motion increases the processing cost of visual perception and as a result of limited processing power in the HVS, considerably reduces visual sensitivity. This phenomenon, referred to as motion suppression [30], happens mainly in low-attentional areas when motion is different to that in high-attentional areas.

Top-down attention, on the other hand, is driven by higher-level cognitive factors and external influences, such as, semantic information, contextual effects, viewing task, and personal preference, expectations, experience and emotions. Top-down attention is slower, longer lasting (sustained), and unlike bottom-up attention, it requires a voluntary effort to shift the gaze. Top-down attention is considered to have a modulatory effect on bottom-up attention [31]. This is illustrated with regard to Figure 1. When shown this image, the attention of different observers would be driven to different pencils (bottom-up). However, if given the search task to identify the light blue pencil, the attention would be drawn to the pencil in the bottom

[FIG1] Illustration of the modulatory effect of top-down attention on bottom-up attention (image "coloring pencils" courtesy of [32]).

right corner. It is, however, extremely difficult for observers to ignore transient cues and hence, bottom-up attention is highly dominant in situations where there is a sudden onset of a visual stimulus. This phenomenon occurs independent of the task and is referred to as attentional capture.

BOTTOM-UP MODELS ONLY PERFORM WELL ON VISUAL SCENES THAT DO NOT CONTAIN ANY SEMANTIC INFORMATION OR ANY INTERESTING AND MEANINGFUL OBJECTS, WHICH IS RARELY THE CASE IN NATURAL IMAGE AND VIDEO CONTENT.

COMPUTATIONAL VISUAL ATTENTION MODELING

Computational VA models aim to predict the gaze locations of human observers. Current models are inspired by early works such as the feature integration theory by Treisman and Gelade [33], guided search by Wolfe et al. [34], or neural-based architecture by Koch and Ullman [35]. The latter model especially constituted a theoretical basis for biologically plausible models incorporating low-level characteristics of the HVS known to contribute to VA, such as multiple-scale processing, contrast sensitivity, and center-surround processing. A recent trend in computational saliency modeling is the development of statistical [36], information theoretic [37], and Bayesian approaches [26], [28], [38]. A main strength of these models is a strong mathematical foundation.

BOTTOM-UP MODELING

The majority of models focus on bottom-up mechanisms to predict visually salient locations [37]–[42]. Common traits of these models are a feature extraction stage followed by a not yet well-understood pooling into a final conspicuity map. Different feature combination strategies were investigated in [43]. The best tradeoff between prediction performance and generalization was achieved by nonlinear competition between salient locations followed by summation. An interesting additive feature integration method is proposed in [44]. In addition to the contribution of individual features, coupling factors were derived from psychophysical evidence to account for complex interactions between features.

A recent study [45] compared the saliency prediction performance of 13 bottom-up models. It was found that the maximum rather than the average predicted saliency correlates considerably better with human saliency recordings. Two models based on multiple-scale contrast-based processing were found to perform best in predicting visual saliency. Despite these findings, it is to date not fully understood how different feature dimensions contribute to overall visual saliency. None of the 13 tested models, for instance, accounts for momentary eye fixations and thus, variations in visual resolution on the retina. Taking into account whether a target can be identified from distractors in peripheral vision (known as crowding effect [46]) is of great concern in visual search tasks though.

Peripheral vision is highly sensitive to temporal activities [47], thus enhancing detection and perception of temporal changes across the visual field. Motion is therefore among the most dominant features to attract attention and thus needs to be an integral feature of any VA model in the context of dynamic visual scenes [29], [36], [48]–[50]. The models in [48] and [49] compute spatial and temporal features independently and fuse them in a pooling stage. Assuming that spatial and motion cues are not separable, the nonparametric models in [50] and [36] outperform earlier models by computing spatiotemporal features based on the phase-spectrum and spatiotemporal local steering kernels, respectively. A biologically inspired spatiotemporal saliency model based on a center-surround framework is proposed in [29]. The incorporation of spatiotemporal aspects into VA models is still an open issue, primarily since human perception of dynamic scenes lacks a theoretical foundation, as is available for still images. From a computational modeling viewpoint, one major challenge is to account for the various combinations of static to dynamic egomotion and scene motion in natural video sequences.

Bottom-up models only perform well on visual scenes that do not contain any semantic information or any interesting and meaningful objects, which is rarely the case in natural image and video content. Furthermore, bottom-up models process the visual scene in a local-to-global manner, meaning, that local features are accumulated into global conspicuity maps. According to this strategy, the number of candidate targets can be high and the scanpath prediction is difficult. A more recent holistic approach shows that the gist of a visual scene is perceived preattentively and can therefore already be integrated prior to the first saccade [26].

TOP-DOWN MODELING

Bottom-up and top-down cues need to be fused in a meaningful way to obtain a single focus of attention. Several works have tackled the difficult task of integrating top-down information with bottom-up features [26], [51]–[53]. A Bayesian framework for contextual guidance is proposed in [26], which is based on parallel computation of local saliency and global context features that enhance object and scene change detection. In visual search tasks, prior knowledge about the target is of particular importance as it strongly influences the search performance (see the coloring pencils example in the section "Bottom-Up and Top-Down Mechanisms"). Therefore, the target-relevant region should be excited, the target-irrelevant regions inhibited, or a combination thereof [54]. The performance of the well-known bottom-up model by Itti et al. [39] was improved by taking into account top-down cues to enable visual search. The degree to which these mechanisms contribute to the overall model needs to be adaptive to the current situation, with bottom-up cues dominating in exploratory (free viewing) conditions and top-down cues dominating in visual search tasks.

Independent of the viewing conditions, neither of the two mechanisms should be entirely suppressed.

IN VISUAL SEARCH TASKS, PRIOR KNOWLEDGE ABOUT THE TARGET IS OF PARTICULAR IMPORTANCE AS IT STRONGLY INFLUENCES THE SEARCH PERFORMANCE.

VISUAL ATTENTION FOR QUALITY ASSESSMENT: RECENT ADVANCES

Increased awareness to the strong interaction between VA and quality perception led to a number of computational methods that integrate VA into quality metrics to potentially improve prediction performance. We discuss in the following the most common VA integration methods and review recent advances for image [2]–[9] and video applications [10]–[20].

COMMON VISUAL ATTENTION INTEGRATION METHODS

We categorize the most common VA integration methods as illustrated in Figure 2. In Method 1, the perceptual difference (PD) between a test (T) and reference stimulus (R) is evaluated independently from the natural scene saliency. In a pooling stage, the perceptual difference is then typically weighted using the saliency map (SM), yielding the final quality score (Q). Assuming that distortions alter attention, the saliency difference (SD) between the reference and distorted stimuli can be used instead of or in addition to the natural scene saliency. Models following Method 2 first segment the image or video frames into salient regions (S) and background (B) using natural scene saliency. The perceptual difference is then computed independently on these regions and combined into an overall quality metric using a weighted summation.

IMAGE QUALITY ASSESSMENT

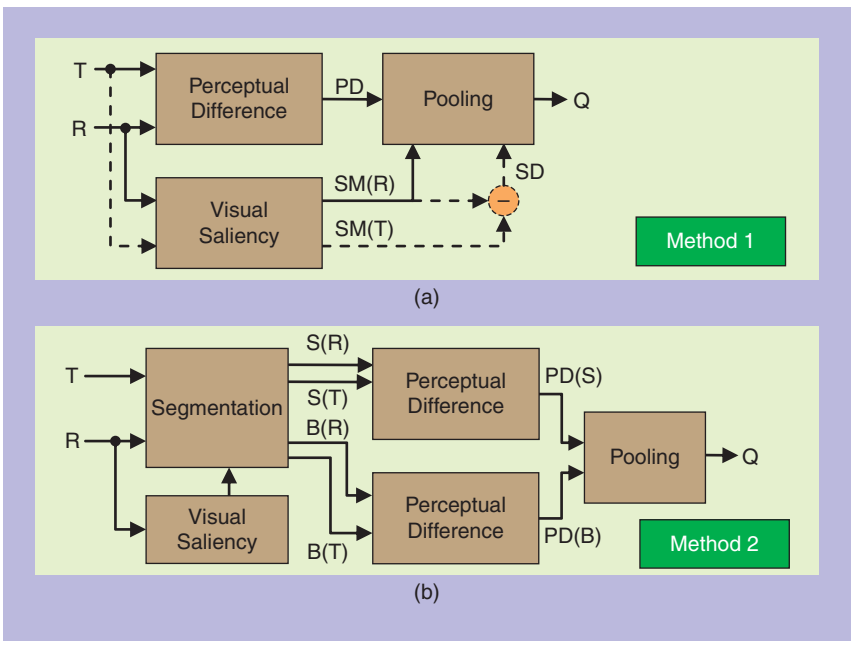
Method 1 is the most widely adopted approach in image quality assessment [2]–[7]. Despite the common integration method, different conclusions arise from these works.

Barland et al. [2] used the Osberger VA model [48] and proposed a multiple-scale VA model for integration into no-reference (NR) blur and ringing metrics for JPEG2000 compressed images. The proposed model yielded a superior performance, which supports the finding in [45] that multiple-scale processing is beneficial for VA models. Sadaka et al. [4] integrated bottom-up saliency [39] into their sharpness metric through multiplicative weighting with the distortion map. The linear correlation coefficient (CC) was enhanced from CC = 0.58 to CC = 0.69. The rather low performance of the original metric, however, provided a big margin for improvement. Moorthy et al. [5] incorporate bottom-up saliency [41] into the structural similarity (SSIM) index [55]. An improvement of CC of approximately 1–4% was achieved across different distortions covered in the test images. No results are reported to validate the statistical significance of the rather low improvements.

Gkioulekas et al. [6] adopt the surprise model in [28] for images and incorporate it into SSIM through weighted summation. The authors found that their surprise model improves SSIM considerably more than the bottom-up model in [39]. It was further found that maximum local saliency provides superior results than averaged local saliency, which is in line with findings in [45], [56]. The improvements to the original SSIM index of approximately 1% in CC are marginal though.

Instead of using a VA model, Ninassi et al. [3] integrated fixation density maps (FDM) from quality-task eye tracking into SSIM and the mean absolute distance (MAD) metric. On the contrary to the other works, no improvements were found in the context of JPEG and JPEG2000 distorted images.

A comprehensive study on incorporating task-free and quality-task eye tracking data into quality metrics (SSIM, visual information fidelity (VIF) [57] criterion, peak signal-to-noise ratio (PSNR), generalized block edge impairment metric (GBIM) [58]) has recently been published by Liu et al. [7]. Statistically significant improvements were found for all metrics, with a superior performance in the case of task-free eye tracking data. The improvement was shown to be larger for images with distinct salient locations, as compared to images that have widely spread saliency. It was further concluded that background distortions should not be neglected, in particular in



[FIG2] **Common methods of integrating VA into quality metrics: (a) Method 1 and (b) Method 2.**

visual scenes where distortion visibility in the background is considerably higher than in the salient region.

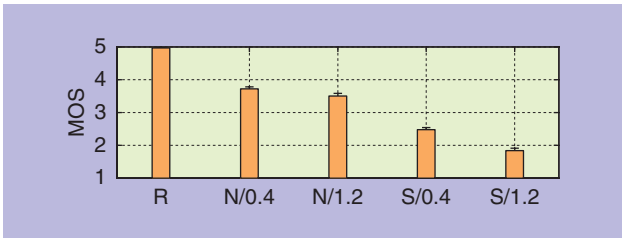
The suitability of Method 2 for VA integration was evaluated in [8], [9]. Larson et al. [8] segmented images into primary regions-of-interest (ROI), secondary ROI, and background based on task-free and quality-task eye tracking data. Five quality metrics [SSIM, VIF, PSNR, visual signal-to-noise ratio (VSNR) [59], and weighted signal-to-noise ratio (WSNR)] were computed independently on these regions and combined using a weighted summation. All metrics received highest weights for the primary ROI, apart from VSNR, which favored the secondary ROI. Superior improvement is reported with task-free rather than the quality-task eye tracking data. Unlike in [7], no improvements were found to be significant.

Engelke et al. [9] proposed an optimization framework for ROI-based image quality metrics in the context of wireless imaging distortions. Significant improvements were found for SSIM, VIF, and PSNR, which are believed to be due to the localized nature of the distortions. Unlike with global distortions that were considered in previous works [3], [5], [8], the impact of the distortion location inside or outside the ROI has a more significant impact. In line with [8], however, it was also found that VSNR received a higher weight for the background, which emphasizes the sensitivity of saliency integration to the distortion measure used.

VIDEO QUALITY ASSESSMENT

Due to the dynamic changes of the visual scene in video applications, it is usually impossible to observe all details within every frame. Our gaze is mainly driven to follow the most salient regions, unlike with images, where sufficiently long viewing times also allow to analyze the background regions. Distortions that occur outside the most salient areas are therefore assumed to have a lower impact on the overall quality and VA concepts can be expected to have a higher impact in video as compared to image applications. The strong attentional guidance due to motion cues and temporal changes plays an essential role and has been implemented in many VA integration methods, as discussed in the following.

Cavallaro et al. [10] integrated a low-level feature-based (motion, color) quality metric with extraction of semantic information by means of face segmentation. Independent quality assessment in the faces and the background, followed by a pooling stage, led to considerable improvements.



[FIG3] MOS for five different distortion classes: (R = reference, S = salient region, N = nonsalient region, 0.4 = 0.4 s distortion length, and 1.2 = 1.2 s distortion length).

Lu et al. [11] modulate the distortion maps of just-noticeable-difference (JND) models as well as PSNR and SSIM using an original VA model based on bottom-up (color, texture, motion) and top-down (faces, skin color) cues. The model accounts for absolute and relative motion as well as motion suppression. All features are pooled using the model in [44]. The JND and quality models were strongly improved.

You et al. [12] integrate top-down cues (faces and text) with the bottom-up model in [39]. Different weighting schemes are tested for VA integration into SSIM and PSNR. Improvements were found only for PSNR but not for SSIM, and it is concluded that SSIM is unsuitable for VA integration. Considering the finding in [7] it may be that this conclusion arises from unsuitable pooling that neglects distortions in the background of the visual scene. In [16], the same group takes into account global quality and motion in addition to local, saliency-based quality analysis. This combination is found to outperform the individual local and global quality measures.

Ma et al. [13] propose a complex VA model to weight spatial distortions without totally neglecting background distortions. In addition, motion suppression is accounted for as well as egomotion of the camera. Integration of the model into SSIM, VIF, and PSNR improved performance of the metrics approximately 8%, 5%, and 3%, respectively.

Engelke et al. [15] conducted a quality-task eye tracking experiment to identify the perceived annoyance of packet loss distortions located either in a salient (S) or nonsalient (N) region. Two different distortion lengths were considered as well (0.4 s and 1.2 s). The mean opinion scores (MOS) presented in Figure 3 reveal that distortions located in salient regions are considerably more annoying than distortions in nonsalient regions. In fact, even the short distortions in the salient regions (S/0.4) received one MOS unit lower than the long distortions in the nonsalient region (N/1.2). Based on these results, a saliency awareness framework for VQM in the context of localized packet loss distortions was proposed [14]. The contemporary temporal trajectory aware VQM (TetraVQM) [60] and PSNR could be improved by penalizing the distortion measures in relation to the underlying content saliency.

Le Meur et al. [17] integrated task-free and quality-task eye tracking data into an original VQM. No improvements were reported with either of the eye tracking data in the context of H.264/AVC compression distortions, which agrees with an earlier study of the same group on images [3]. However, as in the previous study, only simple spatial pooling functions have been considered for VA integration. In a similar study in the context of H.264 compression distortions, Gao et al. [18] report a 4% improvement in CC by integrating a spatiotemporal, bottom-up VA model into SSIM. However, no statistical significance analysis is provided to support the validity of the results. Generally, the global compression distortions considered in these studies can be assumed to have little effect on the VA integration in comparison to, for instance, the localized packet loss distortions considered in [14].

The study by Feng et al. [19] supports the strong impact of localized packet loss distortions in relation to content saliency. Unlike the previously discussed works, this study analyzed the potential benefits of taking into account the saliency difference (SD) between the reference and distorted video (see Method 1 in Figure 2). The intensity, color, and orientation features from the bottom-up model in [39] were extended with a motion model and subject to a weighted summation. By incorporating this model into SSIM, MAD, and the mean squared error (MSE), correlations with subjective quality ratings of up to 0.99 were achieved. Given the relatively large number of seven parameters in the model as compared to the training set of 12 sequences, the model may in fact be overfitted to some degree.

Ćulibrk et al. [20] took a different approach to the two methods depicted in Figure 2. Instead of combining an existing quality metric with a VA model, 35 different features were considered to train a regression tree. Only the five features that had the most significant impact on the metrics performance in the context of MPEG-2 compression distortions were selected. It was found that blocking and blur artifacts were highly annoying in the salient regions whereas temporal distortions were annoying throughout the visual field. The authors concluded that background distortions should not be entirely neglected for successful saliency integration into quality assessment, which supports the conclusions drawn in [7] for images.

SUMMARY OF FINDINGS

From the works discussed in this section, we can summarize that improvement in quality prediction performance due to VA integration is generally superior

1) in video rather than image applications
2) in the case of localized rather than global distortions
3) with task-free instead of quality-task eye tracking data
4) if top-down cues are integrated in addition to bottom-up cues (thus far, usually only faces and text are considered)
5) if motion (relative motion, motion suppression, egomotion) is appropriately integrated in video applications
6) if background distortions are not entirely suppressed but only relative to salient region distortions
7) if multiple-scale analysis is included in the VA model.

Despite many common agreements, some conclusions about the potential benefits of VA integration for quality assessment are controversial. For instance, in [12] it was concluded that SSIM is unsuitable for saliency inclusion whereas [5], [7], and [9] reported particularly good improvements for SSIM. Such controversies are an indication of the many challenges that still need to be solved. Some of the major challenges are identified and discussed in the following section.

CURRENT CHALLENGES

GROUND TRUTH SELECTION

The kind of VA ground truth incorporated into the quality metrics is assumed to have a strong impact on the success of the VA integration performance. Several aspects are of particular interest in this respect.

COMPUTATIONAL MODELS VERSUS PSYCHOPHYSICAL DATA

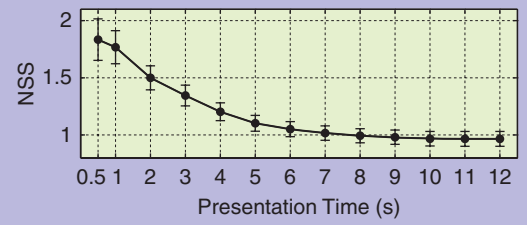
The saliency ground truth used throughout the works discussed in the section "Visual Attention for Quality Assessment: Recent Advances" is either based on computational VA models or psychophysical data. The former has the advantage that it enables automated deployment in image and video processing systems. However, even current state-of-the-art VA models are not reliable predictors of human viewing behavior, often because they focus on bottom-up cues, neglecting important top-down cues and semantic information. Psychophysical experiments, on the other hand, are considered to be a reliable ground truth. To avoid potential modeling errors due to poorly performing VA predictors, we therefore strongly recommend to use psychophysical data as ground truth. As psychophysical experiments find no deployment in real-time applications, it is of great concern to develop VA models that more reliably predict human viewing behavior.

EYE MOVEMENTS VERSUS REGIONS-OF-INTEREST

Eye tracking data is the most common psychophysical ground truth. As eye movements are driven by bottom-up and top-down cues, it is difficult to identify to which degree low-level features and object-level semantics contribute to the resulting FDM. Furthermore, during the search for the most interesting or informative regions, humans do not only attend useful locations [22] and as such, eye tracking recordings do not provide direct insight into which regions are of interest. To obtain more direct insight into perceived interest and its interrelation with eye movements, we conducted an experiment in which human observers hand-labeled ROI in natural images [61]. The resulting ROI maps are compared to FDM from an eye tracking experiment [62] with the aim to identify the presentation time that best predicts the ROI maps. The degree of similarity between FDM and ROI selections is quantified using the normalized scanpath saliency (NSS) [63], as presented in Figure 4. The similarity gradually decreases with presentation time during eye tracking, which suggests that early fixations best predict the ROI. Similar results were reported in [64] and [65]. These findings support the earlier discussion that the gist of a scene is perceived preattentively and thus guides early eye movements (see the section "Bottom-Up Modeling"). The conclusions are expected to be highly task dependent though, since, for instance, a radiologist searching for breast cancer would unlikely attend the target with the early fixations. Similar studies are needed for video, as bottom-up motion cues strongly guide attention and thus might lead to different conclusions.

TASK-FREE VERSUS QUALITY-TASK EYE TRACKING DATA

Whether to use eye tracking data from task-free or quality assessment condition as a ground truth is still an open question. Viewing behavior can change considerably in a visual search task such as quality assessment [51], [66].



[FIG4] NSS between FDM and ROI selections for different presentation times.

Covert VA improves speed and accuracy on many detection, discrimination and localization tasks [23], for which reason observers are sensitized to distortions during quality assessment. This is particularly true since the observer usually has prior knowledge about the distortions (the target). Models that aim to predict gaze patterns recorded under quality assessment task therefore need to be tuned accordingly and attention to distortions needs to be excited relative to the content.

In natural conditions, humans do not view images or video sequences with the aim to identify possible degradations in the content. Their attention is therefore not sensitized to these targets. As the ultimate goal of quality assessment is the prediction of quality perception during these natural conditions, eye tracking data from task-free experiments might in fact be the more sensible choice. The validity of these presumptions is believed to hold particularly for static visual scenes, for images, and is supported by several recent studies [3], [7], [8]. It was found that fixations spread more into the background of the visual scene and thus overestimate the relative impact of distortions in the background to distortions in salient regions [7]. In dynamic visual scenes, on the other hand, fixation durations and locations were not found to be significantly different between task-free and quality assessment conditions [17]. This can be largely explained through the phenomenon of attentional capture due to motion and temporal changes, as discussed in the section "Bottom-Up and Top-Down Mechanisms." In summary, the choice between a task-free and quality assessment task is potentially more crucial in image as compared to video quality assessment. More studies are needed to confirm these observations.

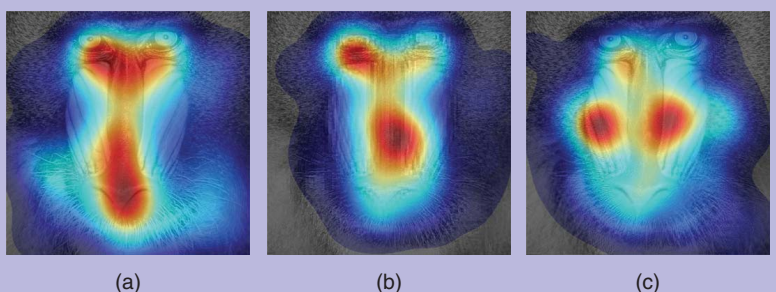
IMPACT OF DISTORTIONS ON VISUAL ATTENTION

The degree to which distortions attract attention in relation to the underlying natural image or video content depends on many influencing factors, such as the natural content saliency and the distortion type, strength, and distribution. In general, distortions that are strongly salient compared to the natural content are expected to attract more attention and thus result in a stronger impact on the overall perceived quality.

GLOBAL VERSUS LOCAL DISTORTIONS

Spatially and spatiotemporally local distortions (e.g., due to packet loss) were shown to attract attention comparably stronger than globally distributed distortions (e.g., due to compression) [19], [67], [68]. This is related to the Bayesian notion of surprise [28], which states that novel events resemble saliency in space and in time and are thus strong attention attractors. High temporal sensitivity in peripheral vision further supports detection of local and time varying distortions. Local distortions therefore alternate gaze patterns relatively strong compared to global distortions. Recent psychophysical evidence supports this rationale. In [69], it was found that global compression distortions do not alter viewing patterns considerably while in [68] it was shown that localized packet loss distortions considerably change viewing behavior.

We studied the shift of gaze patterns during image quality assessment in the case of localized wireless imaging distortions. Figure 5(a) depicts a heat map on the undistorted "Mandrill" image. The distorted versions in Figure 5(b) and (c) exhibit strong blocking distortions and subtle ringing distortions, respectively. Against intuition, the subtle ringing distortions change the gaze pattern considerably more than the strong blocking distortions. Despite the stronger shift, the image in Figure 5(c) received a considerably higher MOS of 64 (on a scale from zero to 100) compared to 25 for the image in Figure 5(b). Covert attention shifts between reference and distorted images thus need to be handled with great caution, as they do not directly relate to quality perception. These observations confirm the earlier discussion that quality-task eye tracking data is unsuitable in case of images, in particular in case of localized distortions.



[FIG5] Heat maps for the image "Mandrill": (a) reference image, (b) image with strong blocking artifacts, and (c) image with subtle ringing artifacts.

DISTORTIONS IN RELATION TO CONTENT SALIENCY

The alternation of viewing behavior was found to be strongly depending on whether distortions are appearing in salient or nonsalient regions [68]. This phenomenon is illustrated for video in Figure 6. The area under the receiver operating characteristic (ROC) curve (AUC) is used to measure the amount of overt attention in the respective distortion regions (salient or nonsalient). As expected, the nonsalient regions are attended less than

the salient regions throughout the video sequence. Upon appearance of the packet loss distortions (A), the gaze is shifted towards the distortions in the nonsalient region, as indicated by the rise in AUC in Figure 6(a). After disappearance (D), the gaze shifts back to the salient region. Unlike in the case of images, we found that the MOS were highly correlated with the AUC in the distortion regions (CC = −0.79). Thus, attention to distortions in video is indeed related to the overall perceived quality, even under quality-task condition.

BOTTOM-UP AND TOP-DOWN CUES INFLUENCE VIEWING BEHAVIOR IN ANY CONTEXT AND SHOULD THEREFORE NOT BE SUPPRESSED ENTIRELY.

The attention shift, however, was not observed for all sequences, as indicated in Figure 6(b). The attention shift appears to be strongly dependent on the relative strength of saliency between content and distortions. The relative location is also important as performance in visual search tasks deteriorates with increased eccentricity in peripheral vision [23]. These findings provide only indications of the complex interaction between content and distortion saliency. Task-free eye tracking experiments on distorted content are needed to study distortion related attention shifts under natural viewing conditions.

VISUAL ATTENTION INTEGRATION

The pooling of the VA and distortion information is probably the most crucial step of VA integration into quality metrics. The psychophysiological mechanisms underlying the interaction between VA and quality perception are not well understood yet. Given the recent psychophysical findings, however, some interesting directions for improved pooling methods can be derived.

SOME GENERAL ISSUES

The combination of model parameters is often done in an ad hoc manner, using simple, Minkowski-like pooling functions. The pooling step typically introduces additional parameters to the model and thus, allows for the designer to better fit the model to the data. A theoretical foundation about pooling methods is needed to comprehend to what degree the improvement is due to saliency integration or due to increased degrees of freedom alone.

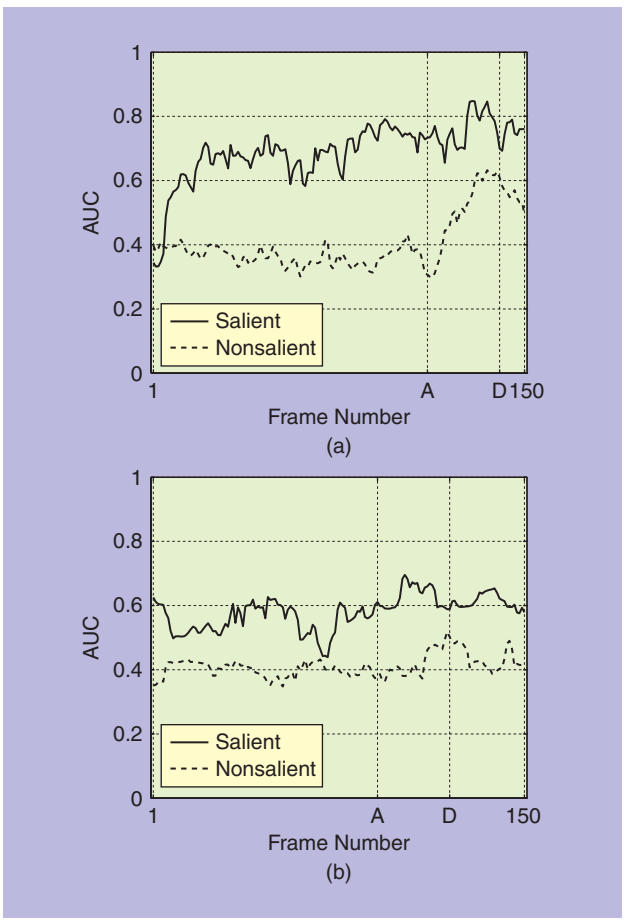
Many studies performed are still using purely bottom-up VA models, even though they are known to not perform well in complex natural scenes. Top-down models therefore need to be included into the pooling stage. Additive rather than multiplicative pooling should be used [54], since both bottom-up and top-down cues influence viewing behavior in any context and should therefore not be suppressed entirely.

SPATIAL POOLING

Image distortions have perceptual impact whether they are in the salient region or not, especially in the context of local distortions. Typical pooling steps, which multiply saliency maps with distortion maps, often suppress background distortions entirely. Depending on the masking properties of the image, background distortions can be strong attractors of attention and are perceived as highly annoying [7]. The pooling method in [13] accounts for background distortions and might constitute a good basis to exploit appropriate pooling of salient region and background distortions.

SPATIOTEMPORAL POOLING

Motion and temporal changes in video have a substantial impact on distortion perception. Due to motion suppression, detection and perception of distortions are considerably reduced in peripheral vision. Spatiotemporal contrast sensitivity functions used in video quality models should therefore be adapted in relation to the motion observed in the visual scene. Attentional capture, on the other hand, counteracts this phenomenon, causing easy detection of spatiotemporally local distortions in the background. Spatiotemporal distortions in the peripheral visual field should therefore not be entirely neglected. Taken these phenomena into account conjointly constitutes a great challenge and requires more sophisticated spatiotemporal VA models.



[FIG6] AUC of two video sequences for (a) a strong and (b) a weak attention shift towards the distortions.

TOWARD MORE APPROPRIATE POOLING METHODS

In most works reviewed in the section "Visual Attention for Quality Assessment: Recent Advances," perceptual distortions and visual saliency are evaluated independently and combined in a pooling stage (see Figure 2). The strong interaction between VA to natural content and distortions, however, might call for more integrated methods that take into account content and distortion saliency simultaneously. In addition, other image and video properties, such as masking effects, need to be accounted for conjointly. To fully understand these interactions and to develop them into advanced pooling techniques, theoretical foundations need to be established first and more psychophysical evidence is needed. Some advances on bottom-up feature integration have been reported in the vision science community [43], [44]. The model in [44] takes into account feature interactions and might thus constitute a suitable candidate for the pooling stage. However, these experiments were carried out on simple stimuli and similar studies are needed for static and dynamic natural scene content.

CONCLUSIONS AND FUTURE DIRECTIONS

The current state of research discussed in this article suggests that there is indeed a benefit of integrating VA into perceptual quality assessment. Most notably, VQM for the assessment of localized artifacts may benefit from the incorporation of VA. However, the existing methods are strongly engineering inspired and the interaction between VA and quality perception is often simplified. Closer collaboration between the image processing and vision science communities is imperative to further enhance this immature field of research.

The following exciting issues were outside the scope of this article but are worth exploring. Most of the works discussed here were based on full-reference quality assessment. VA models are designed to work without any reference and may therefore provide valuable guidance to further develop no-reference quality metrics. Gaze patterns from eye tracking experiments are known to reflect predominantly overt VA. Psychophysiological data, such as through electroencephalography, needs to be investigated to obtain a better understanding of covert VA to distortions in natural content. In the context of multimedia, VA is driven not only by visual cues. Auditory cues are known to be a strong attractor of VA and their impact on attention deployment needs to be explored [70]. Upcoming three-dimensional (3-D) applications constitute an exciting research direction, since additional 3-D cues influence the attention of an observer. These applications induce their own range of distortions, each of them attracting attention to a certain degree that has yet to be investigated.

AUTHORS

Ulrich Engelke (ulrichengelke@gmail.com) received the Dipl.-Ing. degree in electrical engineering in 2004 from RWTH Aachen University, Germany, and the Ph.D. degree in telecommunications in 2010 from the Blekinge Institute of Technology, Sweden. His Ph.D. studies were largely funded by a five-year scholarship awarded through the Royal Institute of Technology (KTH), Sweden. In 2011, he pursued a postdoc position at the University of Nantes, France. Currently he is with the Visual Experiences Group at Philips Research, The Netherlands, working with perception in lighting applications. His research interests include visual scene understanding, human perception, psychophysical experimentation, and signal processing.

Hagen Kaprykowsky (hagen.kaprykowsky@gmail.com) received a German-French double-diploma degree in electrical engineering from the University of Karlsruhe (TH) and INP Grenoble in 2005. Within his diploma thesis, he developed a globally optimal dynamic time-warping algorithm for musical alignment at the IRCAM Centre Pompidou in Paris. He gained his first professional experiences at the German Research Center for Artificial Intelligence in Kaiserslautern, where he worked on the development of adaptive statistical methods for optical character recognition. In 2008 he joined the Image Processing Department of the Fraunhofer Heinrich Hertz Institute. His research interests include perceptual quality assessment and perception-oriented video coding.

Hans-Jürgen Zepernick (hans-jurgen.zepernick@bth.se) received the Dipl.-Ing. degree from the University of Siegen in 1987 and the Dr.-Ing. degree from the University of Hagen in 1994. From 1987 to 1989, he was with Siemens AG, Germany. He is currently a professor of radio communications at the Blekinge Institute of Technology, Sweden. His previous positions include professor of wireless communications at Curtin University of Technology; deputy director of the Australian Telecommunications Research Institute; and associate director of the Australian Telecommunications Cooperative Research Centre. His research interests include radio transmission techniques, mobile multimedia communications, and perceptual quality assessment.

Patrick Ndjiki-Nya (patrick.ndjiki-nya@hhi.fraunhofer.de) received the Dipl.-Ing. in 1997 and the Dr.-Ing. degree in 2008 from the Technische Universität Berlin. From 1997 to 1998, he was involved in the development of a flight simulator at Daimler-Benz. From 1998 to 2001 he was employed as development engineer at DSPecialists. During the same period, he researched content-based video features at Fraunhofer Heinrich Hertz Institute with the purpose of implementation in DSPecialists' DSP solutions. Since 2001, he has been with Fraunhofer Heinrich Hertz Institute, where he was a project manager initially and senior project manager since 2004. He was appointed group manager in 2010. He is a Member of the IEEE.

REFERENCES

[1] J. Wolfe, "Visual attention," in *Seeing,* K. K. D. Valois, Ed. San Diego, CA: Academic, 2000, pp. 335–386.

[2] R. Barland and A. Saadane, "Blind quality metric using a perceptual importance map for JPEG-2000 compressed images," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2006, pp. 2941–2944.

[3] A. Ninassi, O. L. Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2007, vol. 2, pp. 169–172.

[4] N. G. Sadaka, L. J. Karam, R. Ferzli, and G. P. Abousleman, "A no reference perceptual image sharpness metric based on saliency weighted foveal pooling," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2008, pp. 369–372.

[5] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Select. Topics Signal Processing*, vol. 3, no. 2, pp. 193–201, 2009.

[6] I. Gkioulekas, G. Evangelopoulos, and P. Maragos, "Spatial Bayesian surprise for image saliency and quality assessment," in *Proc. IEEE Int. Conf. Image Processing*, Sept. 2010, pp. 1081–1084.

[7] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 971–982, July 2011.

[8] E. C. Larson, C. Vu, and D. M. Chandler, "Can visual fixation patterns improve image fidelity assessment?" in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2008, pp. 2572–2575.

[9] U. Engelke and H.-J. Zepernick, "A framework for optimal region of interest-based quality assessment in wireless imaging," *J. Electron. Imaging (Special Section on Image Quality)*, vol. 19, no. 1, 2010, pp. 1–13.

[10] A. Cavallaro and S. Winkler, "Segmentation-driven perceptual quality metrics," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 2004, vol. 5, pp. 3543–3546.

[11] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory after effects on visual sensitivity and quality evaluation," *IEEE Trans. Image Processing*, vol. 14, no. 11, pp. 1928–1942, 2005.

[12] J. You, A. Perkis, M. Hannuksela, and M. Gabbouj, "Perceptual quality assessment based on visual attention analysis," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2009, pp. 561–564.

[13] Q. Ma, L. Zhang, and B. Wang, "New strategy for image and video quality assessment," *J. Electron. Imaging (Special Section on Image Quality)*, vol. 19, no. 1, 2010, pp. 1–14.

[14] U. Engelke, M. Barkowsky, P. Le Callet, and H.-J. Zepernick, "Modelling saliency awareness for objective video quality assessment," in *Proc. Int. Workshop Quality Multimedia Experience*, June 2010, pp. 212–217.

[15] U. Engelke, R. Pepion, P. Le Callet, and H.-J. Zepernick, "Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss," in *Proc. SPIE/IEEE Int. Conf. Visual Communications and Image Processing*, July 2010.

[16] J. You, J. Korhonen, and A. Perkis, "Attention modeling for video quality assessment: Balancing global quality and local quality," in *Proc. IEEE Int. Conf. Multimedia and Expo*, July 2010, pp. 914–919.

[17] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Process. Image Commun.*, vol. 25, no. 7, pp. 547–558, 2010.

[18] X. Gao, N. Liu, W. Lu, D. Tao, and X. Li, "Spatio-temporal salience based video quality assessment," in *Proc. IEEE Int. Conf. Systems, Man and Cypernetics*, Oct. 2010, pp. 1501–1505.

[19] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency inspired full reference quality metrics for packet-loss-impaired video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 81–88, Mar. 2011.

[20] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokrić, V. Crnojević, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 4, pp. 948–958, Apr. 2011.

[21] D. Burr, M. C. Morrone, and J. Ross, "Selective suppression of the magnocellular visual pathway during saccadic eye movements," *Nature*, vol. 371, no. 6497, pp. 511–513, 1994.

[22] E. Kowler, "Eye movements: The past 25 years," *Vision Res.*, vol. 51, no. 13, pp. 1457–1483, 2011.

[23] M. Carrasco, "Visual attention: The past 25 years," *Vision Res.,* vol. 51, no. 13, pp. 1484–1525, 2011.

[24] A. T. Smith, K. D. Singh, and M. W. Greenlee, "Attentional suppression of activity in the human visual cortex," *Neuroreport,* vol. 11, no. 2, pp. 271–278, 2000.

[25] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev. Neurosci.*, vol. 5, pp. 1–7, June 2004.

[26] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.

[27] B. Khurana and E. Kowler, "Shared attentional control of smooth eye movement and perception," *Vision Res.*, vol. 27, no. 9, pp. 1603–1618, 1987.

[28] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Res.*, vol. 49, no. 10, pp. 1295–1306, 2009.

[29] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[30] B. J. Murphy, "Pattern thresholds for moving and stationary gratings during smooth eye movement," *Vision Res.*, vol. 18, no. 5, pp. 521–530, 1978.

[31] S. Treue, "Visual attention: The where, what, how and why of saliency," *Curr. Opin. Neurobiol.*, vol. 13, no. 4, pp. 428–432, 2003.

[32] M. Maggs. (2007). Colouring pencils. [Online]. Available: http://commons.wikimedia.org/wiki/File:Colouring pencils.jpg

[33] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[34] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: An alternative to the feature integration model for visual search," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 15, no. 3, pp. 419–433, 1989.

[35] C. Koch and S. Ullman, "Shifts in selection in visual attention: Towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[36] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12:15, pp. 1–27, 2009.

[37] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3:5, pp. 1–24, 2009.

[38] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7:32, pp. 1–20, 2008.

[39] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[40] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom–up visual attention," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 5, pp. 802–817, 2006.

[41] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans. Image Processing*, vol. 17, no. 4, pp. 564–573, 2008.

[42] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, June 2009, pp. 1597–1604.

[43] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electron. Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[44] H. C. Nothdurft, "Salience from feature contrast: Additivity across dimensions," *Vision Res.*, vol. 40, no. 10–12, pp. 1183–1201, 2000.

[45] A. Toet, "Computational versus psychophysical bottom–up image saliency: A comparative evaluation study," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 99, 2011.

[46] D. M. Levi, "Crowding—An essential bottleneck for object recognition: A mini review," *Vision Res.*, vol. 48, no. 5, pp. 635–654, 2008.

[47] S. P. McKee and K. Nakayama, "The detection of motion in the peripheral visual field," *Vision Res.*, vol. 24, no. 1, pp. 25–32, 1984.

[48] W. Osberger and A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *Proc. IS&T/SPIE Human Vision and Electronic Imaging VI*, Jan. 2001, vol. 4299, pp. 361–372.

[49] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2006, pp. 815–824.

[50] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[51] F. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Res.*, vol. 45, no. 2, pp. 205–231, 2005.

[52] Y. F. Ma, X. S. Hua, L. Lu, and H. J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[53] C. Kanan, M. H. Tong, L. Zhang, and G. W. Cottrell, "SUN: Top–down saliency using natural statistics," *Vis. Cogn.*, vol. 17, no. 6, pp. 979–1003, 2009.

[54] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.*, vol. 7, no. 1, pp. 1–46, 2010.

[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[56] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.

[57] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Processing*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[58] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Lett.*, vol. 4, no. 11, pp. 317–320, Nov. 1997.

[59] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Processing*, vol. 16, no. 9, pp. 2284–2298, Sept. 2007.

[60] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Select. Topics Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[61] U. Engelke and H.-J. Zepernick, "Psychophysical assessment of perceived interest in natural images: The ROI-D database," in *Proc. SPIE/IEEE Int. Conf. Visual Communications and Image Processing*, Dec. 2011.

[62] U. Engelke, A. J. Maeder, and H.-J. Zepernick, "Visual attention modelling for subjective image quality databases," in *Proc. IEEE Int. Workshop Multimedia Signal Processing*, Oct. 2009, pp. 1–6.

[63] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom–up gaze allocation in natural images," *Vision Res.*, vol. 45, no. 18, pp. 2397–2416, Aug. 2005.

[64] C. M. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur, "Everyone knows what is interesting: Salient locations which should be fixated," *J. Vision*, vol. 9, no. 11:25, pp. 1–22, 2009.

[65] J. Wang, D. M. Chandler, and P. Le Callet, "Quantifying the relationship between visual salience and visual importance," in *Proc. IS&T/SPIE Human Vision and Electronic Imaging XV*, Jan. 2010, vol. 7527.

[66] M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *J. Vision*, vol. 9, no. 3:6, pp. 1–15, 2009.

[67] C. T. Vu, E. C. Larson, and D. M. Chandler, "Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience," in *Proc. IEEE Southwest Symp. Image Analysis and Interpretation*, Mar. 2008, pp. 73–76.

[68] U. Engelke, "Modelling perceptual quality and visual saliency for image and video communications," *Ph.D. dissertation, Blekinge Inst. Technol., Karlskrona*, Sweden, 2010.

[69] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Do video coding impairments disturb the visual attention deployment?" *Signal Process. Image Commun.*, vol. 25, no. 8, pp. 597–609, 2010.

[70] J. S. Lee, F. de Simone, and T. Ebrahimi, "Influence of audio-visual attention on perceived quality of standard definition multimedia content," in *Proc. Int. Workshop Quality Multimedia Experience*, July 2009, pp. 13–18.

[SP]

Margaret H. Pinson, William Ingram, and Arthur Webster

Audiovisual Quality Components

An analysis

The perceived quality of an audiovisual sequence is heavily influenced by both the quality of the audio and the quality of the video. The question then arises as to the relative importance of each factor and whether a regression model predicting audiovisual quality can be devised that is generally applicable.

INTRODUCTION

This article analyzes subjective experiments that explore the relationship between audio quality and video quality, measured separately, and the overall quality of an audiovisual experience. This topic has been explored by at least 12 previous experiments [1]–[11] and one additional experiment described here. Each of these experiments produced a model that mapped audio quality (***a***) and video quality (***v***) to the overall audiovisual quality (***av***). These models are all linear; however, the terms and the values of the coefficients used in the model differ from one experiment to the next.

The goal of this analysis is to describe a flexible audiovisual model that can be applied to a wide range of impairments, applications, source material, and video resolutions. Due to practical limitations, any one subjective experiment can only explore a limited portion of this larger problem. Thus, each experiment contributes to the learning experience: some insights into audiovisual quality, knowledge of what worked well with this experiment, and feedback as to what should be changed for future experiments.

This article contains two main sections. The first describes an experiment conducted at the Institute of Telecommunication Sciences (ITS) in 2010. The second summarizes this and previous experiments, then jointly analyzes the data from all of the summarized experiments to see what conclusions may be

Digital Object Identifier 10.1109/MSP.2011.942470
Date of publication: 1 November 2011

reached. These analyses focus on relative audiovisual qualities (e.g., the quality of one presentation compared to that of another). While important, context effects from specific applications, devices and environments are not considered here. Other studies address context effects (e.g., [17]).

THE GOAL OF THIS ANALYSIS IS TO DESCRIBE A FLEXIBLE AUDIOVISUAL MODEL THAT CAN BE APPLIED TO A WIDE RANGE OF IMPAIRMENTS, APPLICATIONS, SOURCE MATERIAL, AND VIDEO RESOLUTIONS.

ITS 2010 AUDIOVISUAL EXPERIMENT

MOTIVATIONS FOR EXPERIMENT DESIGN

The experiment described in this article was designed to replicate the experiment performed by ITS in 2009 [11] while addressing that experiment's two primary flaws and using high-definition TV (HDTV). The samples for both the ITS 2009 experiment and this, ITS 2010, consisted of a set of audiovisual sequences, where the audio and video were impaired separately. These separate impairments were then combined in all combinations, such that subjects were presented with a full matrix (i.e., all audio impairments combined with all video impairments). This full matrix allows for interesting analysis of variance (ANOVA) on the audiovisual data, to separate the relative impact of audio and video within the overall audiovisual quality scores.

Analysis of ITS 2009 (performed using common intermediate format (CIF) video, 352 × 288) indicated that the video impairments spanned a much wider range of quality than the audio impairments. The question thus arose as to whether the greater weight on video in the audiovisual model was unduly influenced by this unequal distribution. Thus, the ITS 2010 study examined the hypothesis that, if the audio quality spanned nearly the same range as the video quality, then the audio and video quality would be equally important in the overall audiovisual quality.

The ITS 1998 experiment [4] took a different approach. It chose audio impairments that naturally and logically matched with video impairments. This constrained the range of audio and video impairments to those seen in common usage. Since audio requires much lower bitrates than video for CIF resolution and above, the audio quality impairments likewise spanned a limited range of quality. Therefore, audio and video impairments for the ITS 2010 experiment were selected to span approximately the same range of quality.

Another flaw seen in ITS 2009 concerned the types of audio samples used. That experiment intentionally contained 50% sequences with audio consisting of a single person talking. This choice was made out of respect for the extensive research efforts previously conducted using objective models that measure the quality of audio containing a single person talking with no background noise. From an audio compression standpoint, a single person talking is extremely easy to code. Thus, relatively little new information was learned when making comparisons within these audiovisual sequences.

Of the ten audiovisual sequences examined by ITS 2009, only three were associated with an interesting balance such that both the audio quality and the video quality significantly impacted the overall audiovisual quality. For the other seven sequences, the video quality dominated (i.e., video quality explained 89–100% of the distribution of the variance in the subjective data). The three more interesting audio samples contained soft guitar music, crowd noise with an announcer, and music with some talking. In contrast, the ITS 2010 experiment was designed to contain audio that minimizes single person talking samples, and instead emphasizes more complicated audio (e.g., a single person talking with music in the background). A variety of different music types were included, in the hopes that different instruments might elicit different weightings of audio and video quality, thus better representing a wide range of all types of audio.

The decision to use HDTV was motivated by the increasing importance of higher-resolution video. As previous ITS experiments had examined only lower resolution video, examining HDTV would add value.

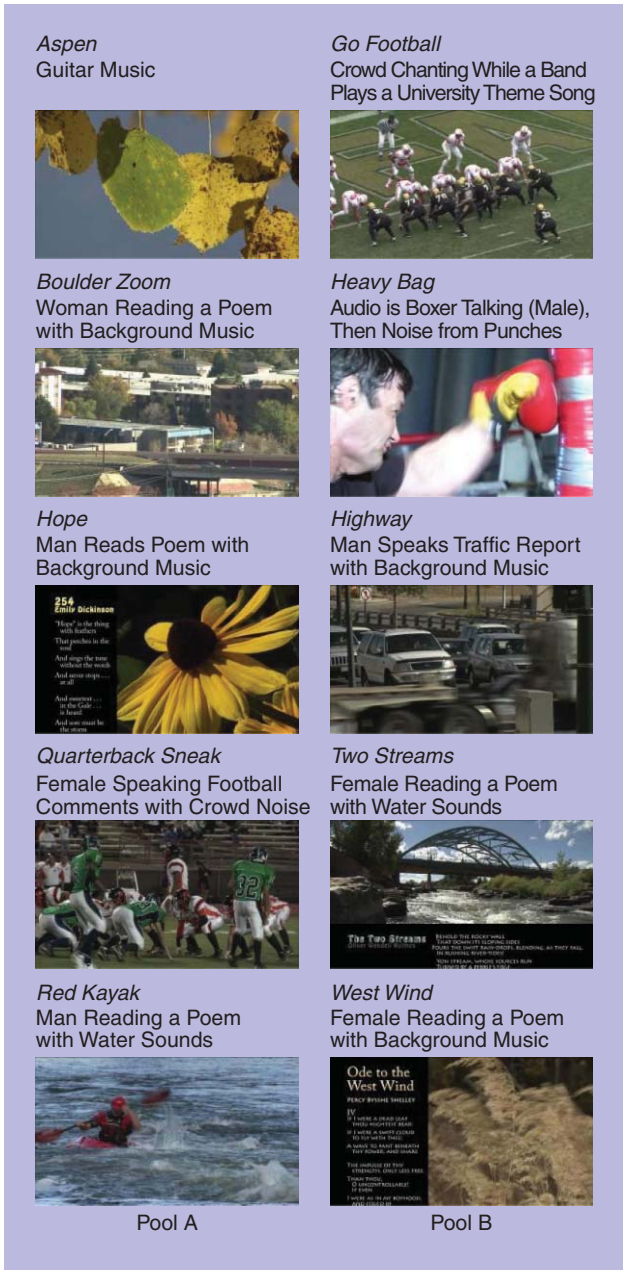
AUDIOVISUAL SEQUENCES

This experiment contained ten audiovisual sequences of 15 s each. Each original sequence contained video that, when visually inspected by an expert, appeared to be of good or better quality. These video sequences were carefully chosen to span a wide range of coding difficulty and a variety of visual characteristics (e.g., scrolling text, fast motion, rapid scene cuts, and random motion). The audio associated with the video sequences was, in most cases, dubbed after the video was created. For most of the sequences, the audio consisted of a single person speaking in a sound-isolated room, combined with either music or background noise. The audio related meaningfully to the video (e.g., a poem *Ode to the West Wind* was paired with grass blowing in the wind). The music represented a wide range of styles and instruments. The sequences were divided into two pools of five, which contained approximately equivalent video and audio characteristics.

Figure 1 shows a sample frame from each sequence, along with a description of the audio and the division of sequences into Pool A and Pool B. These videos were filmed in either 1080p30 or 1080i60 (that is, 1920 × 1080 in either progressive 29.97 frames per second (fps) or interlaced 59.94 fields per second). The audio was recorded in either stereo or mono, at 48 kHz. These audiovisual sequences can be viewed in the Consumer Digital Video Library (www.cdvl.org). They are available for research purposes from that site.

HYPOTHETICAL REFERENCE CIRCUITS

One of the design goals was to produce audio and video hypothetical reference circuits (HRCs) such that each set of



[FIG1] Sample frame from each video sequence displayed below a brief description of the audio.

impairments spanned the full range of quality, from excellent to bad. This experiment included the original video plus six video impairments, and the original audio plus three audio impairments. The audio impairments were Advanced Audio Coding (AAC) at 16, 32, and 48 kb/s. The video impairments were Advanced Video Coding (AVC) at 2, 3.5, and 6 Mb/s, and MPEG-2 at 6, 8.5, and 12 Mb/s. (AAC is also known as ITU-T H.264 [12] and ISO/IEC 14496-10 [13].) The video impairments were chosen from those used in [14], as this provided precise feedback on the expected mean opinion scores (MOS). Reference [14] found that AVC at 3.5 Mb/s was statistically equivalent to MPEG 2 at 8.5 Mb/s. Both received an average MOS between fair and good on the Absolute Category Rating (ACR) scale; see [14, Fig. 3].

Audio impairments were chosen to approximately match the video impairments' range of quality, at the expense of realism. Note that the audio bit rates commonly associated with these video HRCs is high enough that even the lowest associated audio bit rate would cause little impairment to the audio. The selection of the audio and video impairments to span the range of quality was made by several video quality measurement engineers and based on subjective test results from previous experiments (e.g. [16]).

The three AVC encodings were paired with the five video sequences in Pool A (Aspen, Boulder Zoom, Hope, Quarterback Sneak, and Red Kayak). The three MPEG-2 encodings were paired with the five video sequences from Pool B (Go Football, Highway, Heavy Bag, Two Streams, and West Wind). The same three AAC audio impairments were paired with all sequences in both Pool A and Pool B. The pool designations serve only to keep track of which scenes were compressed with AVC and which with MPEG-2.

For each pool, a full matrix of audiovisual impairments was created: three audio impairments plus the original audio, by three video impairments plus the original video, for a total of 16 processed sequences. Each audio impairment was also included in isolation (i.e., audio only, four versions for each of ten sequences) and each video impairment was included in isolation (i.e., video only, four versions for each of ten sequences). Thus, the entire experiment included 240 samples: 160 audiovisual samples, 40 audio only, and 40 video only.

PRESENTATION TO SUBJECTS

Each subject observed and rated half of these sequences, drawn randomly from both Pool A and Pool B, using the ACR scale. Sequences were presented in a random order. The sessions were recorded to Blu-Ray discs. Three randomized splits of the data were created, such that each split included half of the original video sequences (i.e., half of the audiovisual sequences with original audio and original video; half of the original audio only sequences, and half of the original video only sequences). Each split was further broken into three subsessions A, B, and C each containing a random third of the sequences on the disc.

Each Blu-Ray disc contained the same training session, and therefore each subject saw the same training session prior to the test. Each subject saw the three subsessions specific to that disc (A, B, and C) in a random order (e.g., ABC, CAB, BCA), such that no more than three subjects were shown any particular order. Thus, in total there were six Blu-Ray discs each with subsessions A, B, and C specific to that disc:

- Split 1, Even
- Split 1, Odd
- Split 2, Even
- Split 2, Odd
- Split 3, Even
- Split 3, Odd.

Each pair of discs (even plus odd) contained all video sequences. The subjects watched the video on a TV-Logic LVM-460WD professional grade 46-in liquid crystal display (LCD) monitor in a sound-isolated room. They heard the audio over NHT Audio LLC speakers; main speaker model A-20 with subwoofer model B-20. A total of 54 naïve subjects ran through the experiment using the Blu-Ray discs. The subjects were primarily students from the local university who responded to an online advertisement. No subjects had participated in a similar test within the previous six months. Each subject's participation was limited to 90 min.

To analyze the reliability of this method, the following analysis was performed. For each of the 240 processed sequences, a comparison was made between the individual data resulting from one Blu-Ray disc, and the MOS for that clip computed from the other five Blu-Ray discs. The comparison was made using the Student's t-test at the 99% confidence level. For example, the Student's t-test computed with 99% confidence whether the individual opinion scores for Clip 1 on disc "Split 1, Even" appeared to be a random sample from a normal distribution with mean equal to the MOS for that clip from the other two discs where it appeared. This test was performed with the raw opinion scores as specified by the subjects (i.e., no scaling was performed).

Of the Student's t-test comparisons, 6% indicated a different mean at the 99% confidence level. This indicates a fair degree of reliability, given the known noise in subjective data and the impact of ordering effects. We expect some differences in the MOS from one viewer ordering to another, which is the motivation for maximizing the number of orderings presented to subjects. Thus, the viewers of all six Blu-Ray discs may reasonably be interpreted as one single set of viewers. (Note: when opinion scores from one disc are compared to another disc, this value rises to from 6% to approximately 10%.)

ITS 2010 ANALYSIS

Statistical analysis of this experiment will be presented in the next section.

SUMMARY OF ALL EXPERIMENTS

OVERVIEW

Due to practical limitations, any one subjective experiment can only explore a limited portion of the larger problem of describing a flexible model for describing audiovisual quality applicable to a wide range of sample types. Thus, each experiment provides a learning experience: some insights into audiovisual quality, knowledge of what worked well with this experiment, and feedback as to what should be changed for future experiments. This section will examine previous experiments as well as the ITS 2010 experiment and see what conclusions can be reached.

Table 1 summarizes the design of prior experiments that generated an equation. Experiments are listed in chronological order, because previous studies may have influenced later experiments. Each of these experiments performed separate subjective testing of the audio quality of audio-only sequences (i.e., no picture), the video quality of video-only sequences (i.e., silent), and the audiovisual quality of sequences containing some combinations of those audio and video samples. Most of the subjective experiments used a single stimulus methodology that measured people's opinions using MOS. While the majority of experiments mentioned in this article used single stimulus and MOS, one used double stimulus and differential MOS (DMOS). To avoid complicated language, the subjective data for that experiment will occasionally be incorrectly referred to as MOS. Where a single paper listed results from multiple experiments, each experiment is listed on a separate line of the table (e.g., the BT study).

Table 1 lists the video impairments, audio impairments, and the number of processed sequences that were used for the audiovisual portion of each experiment only. Additional impairments of video or audio sequences may have been used in the video-only or audio-only subexperiments and their analysis. The number of processed sequences listed includes the originals, if included in that experiment. Unless otherwise specified, the three types of stimuli (audio only, video only, and audiovisual) were presented in separate sessions. Taken together, these experiments span a wide range of applications, video resolutions, video impairments, audio impairments, and source material.

Table 2 summarizes the equations calculated from each experiment (column "Model") and the Pearson correlation of each model (column "ρ"). All models identified in the original papers that predict audiovisual quality using a combination of audio and video quality are shown. The column "Range of MOS" shows the relative range of MOS spanned by audio-only, video-only, and audiovisual subjective data. The column "Type Comparison" lists the Pearson correlation between audio-only MOS and audiovisual MOS as well as the Pearson correlation between video-only MOS and audiovisual MOS. The final column, "Dominant Factor," lists the subjective conclusion reached by the people conducting the experiment regarding the relative contributions of audio and video in the overall audiovisual quality. This information should be understood in the context of the experimental designs summarized in Table 1.

The models identified in Table 2 are numbered by the model type as follows:

$$1: \hat{y} = \alpha + \mu(\boldsymbol{a} \times \boldsymbol{v})$$
$$2: \hat{y} = \alpha + \beta\, \boldsymbol{a} \times \gamma \boldsymbol{v}$$
$$3: \hat{y} = \alpha + \gamma \boldsymbol{v} + \mu\, (\boldsymbol{a} \times \boldsymbol{v})$$
$$4: \hat{y} = \alpha + \beta\, \boldsymbol{a} + \gamma \boldsymbol{v} + \mu(\boldsymbol{a} \times \boldsymbol{v}).$$

The term $\boldsymbol{a}$ represents the measured audio quality MOS; $\boldsymbol{v}$ represents the measured video quality MOS; and $\boldsymbol{av}$ represents the measured audiovisual quality MOS. The term $\hat{y}$ represents the predicted audiovisual quality. Occasionally a paper will mention the accuracy of one of these model types but not the coefficients. In this case, the model coefficients are left Greek variables. France Telecom additionally presented a model that applied a logarithmic transformation to $\boldsymbol{av}$ and thus estimated

[TABLE 1] COMPARISON OF EXPERIMENT DESIGN FROM DIFFERENT LABORATORIES' INVESTIGATIONS INTO AUDIOVISUAL MODELS.

LABORATORY	FOCUS	DESIGN	SIZE	ENVIRONMENT	VIDEO	AUDIO	SCALE
BELLCORE 1993 [1]	ENTERTAINMENT TELEVISION NTSC	FULL MATRIX OF FIVE AUDIO ONLY BY FIVE VIDEO ONLY. ORIGINALS NOT RATED.	TWO ORIGINALS 50 PVS 18-S CLIPS	TELEVISION MONITOR (CRT) SPEAKERS	RANDOM NOISE, SIMULATED VIDEO CONTENT: ■ CONVERSATION WHILE WALKING ON A BUSY STREET (VARIETY AND MOTION) ■ CONVERSATION WHILE SITTING IN A LIBRARY (SOME HEAD-AND-SHOULDERS, LITTLE MOTION)	TEMPORALLY CORRELATED NOISE, SIMULATING A LOW BIT RATE VOICE CODER AUDIO CONTENT: SPEECH, MAY OR MAY NOT HAVE BACKGROUND NOISE (NOT SPECIFIED)	NINE-POINT SCALE, (EXCELLENT, GOOD, FAIR, POOR, UNSATIS-FACTORY) SIMILAR TO ACR
BELLCORE 1994 [2]	ENTERTAINMENT TELEVISION NTSC	IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]	SIMULATED BLURRING FROM A HORIZONTAL FIR LOW-PASS FILTER VIDEO CONTENT IDENTICAL TO BELLCORE 1993 [1]	MODULATED NOISE REFERENCE UNIT (MNRU) [15] AUDIO CONTENT IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]
BELLCORE 1995 [3]	ENTERTAINMENT TELEVISION NTSC	THREE FULL MATRICES, EACH THREE AUDIO ONLY BY THREE VIDEO ONLY	TWO ORIGINALS 54 PVS 18-S CLIPS	IDENTICAL TO BELLCORE 1993 [1]	VIDEO IMPAIRMENTS: ■ SAME AS [1] ■ SAME AS [2] ■ SIMULATED BLOCKINESS VIDEO CONTENT IDENTICAL TO BELLCORE 1993 [1]	AUDIO IMPAIRMENTS: ■ SAME AS [1] ■ SAME AS [2] ■ RANDOM NOISE AUDIO CONTENT IDENTICAL TO BELLCORE 1993 [1]	IDENTICAL TO BELLCORE 1993 [1]
ITS (1998) [4]	VIDEO-TELECON-FERENCE (VTC) NTSC	AUDIOVISUAL IMPAIR-MENTS CREATED, AND THEN SPLIT INTO AUDIO ONLY AND VIDEO ONLY AUDIOVISUAL BIT RATE 128–1,536 kb/S.	SIX ORIGINALS 48 PVS 5–9-S CLIPS	PC MONITOR (CRT) PC SPEAKERS	ANALOG, H.261, AND PROPRIETARY CODER VIDEO CONTENT: VTC (E.G., HEAD-AND-SHOULDERS, PEOPLE AT A TABLE, MAP WITH POINTER)	ANALOG, G.711, G.722, G.728, AND PROPRIETARY CODER AUDIO CONTENT: ■ SIX WITH SPEECH	ACR FIVE-POINT SCALE
FRANCE TELECOM/CNET 1998 [5]	VIDEO-TELECON-FERENCE (VTC) PAL	FULL MATRIX OF FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	TWO ORIGINALS 32 PVS 10-S CLIPS	MONITOR AND LOUDSPEAKERS	ORIGINAL, CIF AND QCIF AT 12 OR 25 FPS FROM 172 TO 456 kb/S VIDEO CONTENT: VIDEOCONFERENCE	ORIGINAL, G.722,G.711, AND G.728 FROM 16 TO 56 kb/S AUDIO CONTENT: ■ TWO WITH SPEECH (ONE MALE, ONE FEMALE)	ACR FIVE-POINT SCALE
KPN RESEARCH 1997 [6], [7]	BROADCAST TELEVISION PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	TWO ORIGINALS 32 PVS 25-S CLIPS	TELEVISION MONITOR (CRT) STEREO LOUDSPEAKERS	SPATIAL FILTERING OF LUMINANCE SIGNAL IN HORIZONTAL DIRECTION. VIDEO CONTENT: ■ COMMERCIALS	CD QUALITY, BAND LIMITED TO WIDE BAND, AM RADIO, TELEPHONE QUALITY CONTENT NOT SPECIFIED.	ACR NINE-POINT SCALE
BT 2004 [8] EXPERIMENT #1	VIDEO-TELECONFERENCE PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	TWO ORIGINALS 32 PVS 5-S CLIPS ORIGINALS WERE RATED	TELEVISION MONITOR (CRT) SPEAKERS	EMULATED BLOCKINESS VIDEO CONTENT: LOW MOTION, LOW COMPLEXITY	MNRU LEVEL THREE TO 24 AUDIO CONTENT: ■ TWO WITH SPEECH (ONE MALE, ONE FEMALE)	DSCQS 100-POINT SCALE
BT 2004 [8] EXPERIMENT #2, LOW COMPLEXITY	VIDEO-TELECONFERENCE PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	ONE ORIGINAL 16 PVS 5-S CLIPS ORIGINALS WERE RATED	TELEVISION MONITOR (CRT) SPEAKERS	EMULATED BLOCKINESS VIDEO CONTENT: HEAD-AND-SHOULDERS	MNRU LEVEL THREE TO 21 AUDIO CONTENT: ■ SPEECH (MALE)	SINGLE STIMULUS (SSQS) FIVE-POINT SCALE
BT 2004 [8] EXPERIMENT #2, HIGH COMPLEXITY	VIDEO-TELECONFERENCE PAL	FULL MATRIX FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	ONE ORIGINAL 16 PVS 5-S CLIPS ORIGINALS WERE RATED	TELEVISION MONITOR (CRT) SPEAKERS	EMULATED BLOCKINESS VIDEO CONTENT: BICYCLE RACE	MNRU LEVEL THREE TO 21 AUDIO CONTENT: ■ SPEECH	SINGLE STIMULUS (SSQS) FIVE-POINT SCALE

IEEE SignalProcessing
Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page

NATIONAL UNIVERSITY OF SINGAPORE AND EPFL (2006) [9]	ENTERTAINMENT OVER 3GPP QCIF	PARTIAL MATRIX OF FOUR AUDIO ONLY BY FOUR VIDEO ONLY.	SIX ORIGINALS 48 PVS ≈8-S ORIGINALS NOT RATED	PC MONITOR HEADPHONES	AVC FROM 24 TO 48 kb/S CODED AT 8 FPS ENTERTAINMENT CONTENT	MPEG-4 AAC-LC (LOW COMPLEXITY) FROM 8 TO 32 kb/S MONO AUDIO AUDIO CONTENTS: ■ ONE WITH SPEECH ■ ONE WITH MUSIC ■ FOUR MIXED	ACR 11-POINT SCALE
DEUTSCHE TELEKOM 2009 [10]	HDTV WITH PACKET LOSS IMPAIRMENT-FACTOR-BASED MODEL	PARTIAL MATRIX OF AUDIO-ONLY AND VIDEO-ONLY IMPAIRMENTS.	FIVE ORIGINALS 245 PVS 16-S CLIPS ORIGINALS WERE RATED.	PROFESSIONAL GRADE MONITOR (LCD) PROFESSIONAL GRADE SPEAKERS	AVC FROM TWO TO 16 Mb/S, WITH PACKET LOSS: FREEZING FROM 0% TO 0.25% AND SLICING FROM 0% TO 4%	MP2 FROM 48 TO 192 kb/S AND AAC AT 48 kb/S, WITH PACKET LOSS FROM 0% TO 8% AUDIO CONTENTS: ■ ONE WITH SPEECH ■ TWO WITH MUSIC ■ TWO MIXED	ACR 11-POINT SCALE, MAPPED TO 100-POINT SCALE
ITS (2009) [11]	ENTERTAINMENT TELEVISION CIF	TWO FULL MATRICES, EACH FOUR AUDIO ONLY BY FOUR VIDEO ONLY	TEN ORIGINALS 160 PVS 11–12-S CLIPS ORIGINALS WERE RATED	PC MONITOR PC SPEAKERS	H.263, VC-1, AVC, AND MPEG-2 FROM 75 TO 800 Kb/S ENTERTAINMENT CONTENT	M3, PCM, AND WMA FROM FOUR TO 32 kb/S MONO AUDIO AUDIO CONTENTS: ■ FIVE WITH SPEECH ■ THREE WITH MUSIC ■ TWO MIXED	ACR FIVE-POINT SCALE
ITS (2010)	ENTERTAINMENT TELEVISION HDTV	TWO FULL MATRICES, EACH FOUR AUDIO-ONLY BY FOUR VIDEO-ONLY SESSIONS HAD A RANDOM MIX OF AUDIO ONLY, VIDEO ONLY, AND AUDIO-VISUAL	TEN ORIGINALS 160 PVS 15-S EACH ORIGINALS WERE RATED	PROFESSIONAL GRADE TELEVISION MONITOR (LCD) PROFESSIONAL GRADE SPEAKERS	AVC FROM TWO TO SIX Mb/S AND MPEG-2 FROM SIX TO 12 Mb/S ENTERTAINMENT CONTENT	AAC FROM 16 TO 48 kb/S PLUS ORIGINAL STEREO AUDIO AUDIO CONTENTS: ■ TWO MUSIC ■ EIGHT SPEECH WITH MUSIC OR BACKGROUND NOISE	ACR FIVE-POINT SCALE

ln($\boldsymbol{av}$) rather than $\boldsymbol{av}$. This improved the Pearson correlation of their model number one from 0.956 to 0.980 [5].

META-ANALYSIS

Caution should be taken when vertically comparing the Pearson correlation (ρ) values in Table 2, because the denominator measures the range of quality within this particular experiment. Thus, while $\rho = 0.80$ would be poor for an experiment that contains a wide range of quality, $\rho = 0.80$ might be excellent for an experiment that spans a very limited range of quality (e.g., only high-quality video sequences suitable for broadcast television). Within each experiment (i.e., one horizontal row of the table), the correlation values and ranges can all be directly compared. (A purist might disagree. In some experiments, different viewer pools were used for audio only, video only, and audiovisual sessions.)

Likewise, when examining the range of MOS values spanned by each experiment, note the range of values available for that particular experiment (i.e., five-, nine-, 11-, or 100-point scale), the number of source sequences used (from one to ten), and the number of processed video sequences (PVSs) (from 16 to 245).

The oldest published studies were conducted by Bellcore from 1993 to 1995. These three separate experiments focused on standard definition television measured with a very small number of sequences (two) and artificial impairments. All three experiments reached the same conclusion: audio and video qualities are equally important in the overall audiovisual quality. Four later studies disagreed with Bellcore's conclusion: ITS 1998, KPN Research, Deutsche Telekom, and ITS 2009. These studies concluded that video quality was more influential than audio quality on the overall audiovisual quality. However, it can be observed that the range of quality spanned by audio is somewhat less than that spanned by video for these four experiments. When comparing max($\boldsymbol{a}$)-min($\boldsymbol{a}$) to max($\boldsymbol{v}$)- min($\boldsymbol{v}$), the audio range is 84%, 67%, 75%, and 47% of the video range, respectively. By contrast, the range of quality spanned by audio is similar to that spanned by video for the three Bellcore experiments (106%, 94%, and 97%, respectively). Thus, it is possible that this difference biased the experiments in favor of video. The three BT studies and the Singapore/EPFL study were inconclusive on this issue.

The ITS 2010 study described earlier was designed to test the hypothesis that, if the audio quality spanned nearly the same range as the video quality, then the audio and video quality would be equally important in the overall audiovisual quality. The range of quality spanned by audio was intended to be identical to that spanned by video, but ended up being slightly larger (113%). The ITS 2010 study agrees with Bellcore, concluding that audio and video have roughly the same influence on the overall audiovisual quality.

Bellcore also concluded that only the cross term ($\boldsymbol{a} \times \boldsymbol{v}$) is needed to predict the overall audiovisual quality. This conclusion is upheld by the generally stellar performance of this model in the other experiments conducted since then (see Model 1 in Table 2). Only two studies disagree with this

[TABLE 2] COMPARISON OF SUBJECTIVE AUDIOVISUAL MODELS FROM DIFFERENT LABORATORIES' EXPERIMENTS.

LABORATORY	MODEL	ρ	RANGE OF MOS	TYPE COMPARISON	DOMINANT FACTOR
BELLCORE 1993 [1]	1: $\hat{y} = 1.295 + 0.1077(\boldsymbol{a} \times \boldsymbol{v})$	0.99	$\boldsymbol{a}$ = [1.0 TO 8.2] $\boldsymbol{v}$ = [1.9 TO 8.7] $\boldsymbol{av}$ = [1.9 TO 8.3]	UNKNOWN	BOTH AUDIO AND VIDEO HAVE ROUGHLY THE SAME INFLUENCE
BELLCORE 1994 [2]	1: $\hat{y} = 1.07 + 0.1106(\boldsymbol{a} \times \boldsymbol{v})$	0.99	$\boldsymbol{a}$ = [1.4 TO 7.4] $\boldsymbol{v}$ = [1.5 TO 7.9] $\boldsymbol{av}$ = [1.8 TO 7.5]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.67 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.68 ρ	BOTH AUDIO AND VIDEO HAVE ROUGHLY THE SAME INFLUENCE
BELLCORE 1995 [3]	1: $\hat{y} = 1.912 + 0.114(\boldsymbol{a} \times \boldsymbol{v})$	0.99	$\boldsymbol{a}$ = [1.2 TO 7.3] $\boldsymbol{v}$ = [1.8 TO 8.1] $\boldsymbol{av}$ = [1.7 TO 7.3]	UNKNOWN	(MODEL CONSISTENT: ESSENTIALLY THE SAME AS [1] AND [2])
ITS (1998) [4]	1: $\hat{y} = 1.514 + 0.121(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -0.677 + 0.217\boldsymbol{a} + 0.888\boldsymbol{v}$ 4: $\hat{y} = 0.517 - 0.0058\boldsymbol{a} + 0.654\boldsymbol{v} + 0.042(\boldsymbol{a} \times \boldsymbol{v})$	0.927 0.978 0.980	$\boldsymbol{a}$ = [1.5 TO 4.6] $\boldsymbol{v}$ = [1.0 TO 4.7] $\boldsymbol{av}$ = [1.1 TO 4.7]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.41 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.97 ρ $\boldsymbol{a}$ AND $\boldsymbol{v}$ = 0.29 ρ	VIDEO QUALITY
FRANCE TELECOM/CNET 1998 [5]	1: $\hat{y} = 1.76 + 0.10(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -0.13 + 0.35\boldsymbol{a} + 0.57\boldsymbol{v}$	0.960 0.956	$\boldsymbol{a}$ = [1.9 TO 4.5] $\boldsymbol{v}$ = [1.4 TO 4.8] $\boldsymbol{av}$ = [1.5 TO 4.9]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.42 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.86 ρ	COMPARED PASSIVE AND CONVERSATIONAL CONTEXT
KPN RESEARCH 1997 [6], [7]	1: $\hat{y} = 1.45 + 0.11(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = \alpha + \beta\,\boldsymbol{a} + \gamma\,\boldsymbol{v}$ 4: $\hat{y} = 1.12 + 0.007\boldsymbol{a} + 0.24\boldsymbol{v} + 0.088(\boldsymbol{a} \times \boldsymbol{v})$	0.97 0.96 0.98	$\boldsymbol{a}$ = [3 TO 7] $\boldsymbol{v}$ = [2 TO 8] $\boldsymbol{av}$ = [2 TO 8]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.33 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.90 ρ	VIDEO QUALITY
BT 2004 [8] EXPERIMENT 1	1: $\hat{y} = \alpha + \mu(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = 4.26 + 0.59\boldsymbol{a} + 0.49\boldsymbol{v}$ 4: $\hat{y} = -3.34 + 0.85\boldsymbol{a} + 0.76\boldsymbol{v} + -0.01\,(\boldsymbol{a} \times \boldsymbol{v})$	0.72 0.97 0.99	$\boldsymbol{a}$ = [0 TO 63] $\boldsymbol{v}$ = [0 TO 71]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.74 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.62 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
BT [8] LOW COMPLEXITY	1: $\hat{y} = 1.15 + 0.17(\boldsymbol{a} \times \boldsymbol{v})$	0.85	$\boldsymbol{a}$ = [1.2 TO 4.8] $\boldsymbol{v}$ = [1.0 TO 4.6]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.61 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.55 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
BT [8] HIGH COMPLEXITY	1: $\hat{y} = \alpha + \mu\,(\boldsymbol{a} \times \boldsymbol{v})$ 3: $\hat{y} = 0.95 + 0.25\,\boldsymbol{v} + 0.15(\boldsymbol{a} \times \boldsymbol{v})$	0.79 0.85	$\boldsymbol{a}$ = [1.2 TO 3.8] $\boldsymbol{v}$ = [1.0 TO 4.3]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.44 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.68 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
NATIONAL UNIVERSITY OF SINGAPORE AND EPFL (2006) [9]	1: $\hat{y} = 1.98 + 0.103\,(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -1.51 + 0.456\boldsymbol{a} + 0.770\boldsymbol{v}$	0.94 0.94	$\boldsymbol{a}$ = [6 TO 9] $\boldsymbol{v}$ = [2 TO 8] $\boldsymbol{av}$ = [2 TO 8]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.55 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.67 ρ	BOTH CONTRIBUTE SIGNIFICANTLY
DEUTSCHE TELEKOM 2009 [10] (SEE NOTE BELOW)	1: $\hat{y} = 30.917 + 0.007\,(\boldsymbol{a} \times \boldsymbol{v})$ 3: $\hat{y} = 27.805 + 0.129\boldsymbol{v} + 0.006\,(\boldsymbol{a} \times \boldsymbol{v})$	0.95 0.96	$\boldsymbol{a}$ = [30 TO 90] $\boldsymbol{v}$ = [20 TO 100] $\boldsymbol{av}$ = [30 TO 90]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.49 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.83 ρ	VIDEO QUALITY
ITS (2009) [11]	1: $\hat{y} = 1.1096 + 0.1959\,(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -0.5875 + 0.3599\boldsymbol{a} + 0.8037\boldsymbol{v}$ 4: $\hat{y} = 0.7500 - 0.0452\boldsymbol{a} + 0.3882\boldsymbol{v} + 0.1250\,(\boldsymbol{a} \times \boldsymbol{v})$	0.93 0.96 0.97	$\boldsymbol{a}$ = [2.3 TO 3.8] $\boldsymbol{v}$ = [1.3 TO 4.5] $\boldsymbol{av}$ = [1.0 TO 4.9]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.34 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.92 ρ	VIDEO QUALITY
ITS (2010)	1: $\hat{y} = 0.9616 + 0.1919\,(\boldsymbol{a} \times \boldsymbol{v})$ 2: $\hat{y} = -1.2757 + 0.6304\boldsymbol{a} + 0.6807\boldsymbol{v}$ 4: $\hat{y} = 0.9845 - 0.0525\boldsymbol{a} + 0.0274\boldsymbol{v} + 0.1969\,(\boldsymbol{a} \times \boldsymbol{v})$	0.96 0.94 0.96	$\boldsymbol{a}$ = [1.1 TO 4.6] $\boldsymbol{v}$ = [1.6 TO 4.7] $\boldsymbol{av}$ = [1.3 TO 4.8]	$\boldsymbol{a}$ AND $\boldsymbol{av}$ = 0.68 ρ $\boldsymbol{v}$ AND $\boldsymbol{av}$ = 0.66 ρ	BOTH AUDIO AND VIDEO HAVE ROUGHLY THE SAME INFLUENCE

NOTE: Information for Deutsche Telekom Model 1 was received in a private correspondence from Marie–Neige Garcia of Deutsche Telekom.

conclusion. The first is BT 2004, which shows a significant reduction in model accuracy when moving from the best model presented (additive) to the multiplicative model; and the second is BT High Complexity, which shows a moderate drop in correlation. It is possible that the very small number of video sequences used in these studies (two and one video sequences, respectively) resulted in measurement inaccuracies. Even so, BT concludes that people integrate audio and video errors together using a multiplicative rule, and that the true formula depends upon context and the test material under consideration [8]. Thus, in general, we see only a small drop in correlation when moving from the ideal model to the multiplicative model. The ITS 2010 study confirms this robust behavior of a model containing only the cross term.

While the other types of models (Models 2–4) have generally good performance, there is little agreement from one experiment to the next concerning the relative weight that should be assigned to β, γ, and μ. Some of these weights are very different indeed. For example, compare γ and μ for Model 2 as computed by ITS 1998 and BT Experiment 1. Likewise, there is no agreement as to which of these models is best (Models 2–4). This is problematic for someone who wishes to apply one of the other types of models, since it is unclear which set of weights should be chosen. The practical problem with the theory presented by BT (i.e., that different applications drive these differences) is that we do not have sufficient information available to say with any confidence the exact form of that model or the exact weights that are most appropriate for specific use case scenarios. Moreover, the accuracy gain for using one of the other models appears to be insignificant when the subjective experiment is designed with an approximately equal range of audio and video (compare Models 1–4 for ITS 2010).

CONCLUSIONS

There is no apparent pattern of relationship between the accuracy of the multiplicative model and the authors' conclusions as

to the dominant factor (audio quality or video quality) on the overall audiovisual quality. This and the previous analyses (presented earlier) indicate that audio quality and video quality are equally important in the overall audiovisual quality. The application drives the range of audio quality and video quality examined and thus produces the appearance that one factor has greater influence than the other. The underlying perceptual model is invariant to application.

AUDIO QUALITY AND VIDEO QUALITY ARE EQUALLY IMPORTANT IN THE OVERALL AUDIOVISUAL QUALITY.

The most important overall conclusion is that only the cross term ($a \times v$) is needed to predict the overall audiovisual quality. It provides us with a simple and reasonably accurate model that has been tested in a wide variety of circumstances, from CIF to HDTV, from video conferencing to broadcast television, both coding only and with transmission errors, in a professional viewing/listening environment and on a PC. One missing factor is the impact of audiovisual synchronization errors (e.g., lip synchronization) on audiovisual quality. While many studies have been undertaken on audiovisual synchronization, further work is ongoing. A preliminary investigation on this topic undertaken by our lab is available in [16].

ACKNOWLEDGMENT

The ITS experiments presented herein were only possible due to contributions of video quality expertise from Stephen Wolf, audio quality expertise from Stephen Voran and Andrew Catellier, and computer programming from Scott Hanes.

Certain commercial equipment, materials, and/or programs are identified in this report to specify adequately the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration (NTIA), nor does it imply that the program or equipment identified is necessarily the best available for this application.

AUTHORS

Margaret H. Pinson (Margaret@its.bldrdoc.gov) received her B.S. and M.S. degrees in computer science from the University of Colorado at Boulder, in 1988 and 1990, respectively. Since 1988, she has been working as a computer engineer at ITS, an office of the NTIA in Boulder, Colorado. Her goal is to develop automated metrics for assessing the performance of video systems and actively transfer this technology to end users, standards bodies, and U.S. industry.

William Ingram (Bing@its.bldrdoc.gov) earned his B.S. and M.S. degrees in electrical engineering from Oklahoma State University in 1980 and 1981, respectively. Before joining NTIA/ITS in the late 1980s, he worked as an electrical engineer for the United States Department of Interior, Bureau of Reclamation. He has worked on a wide variety of projects at ITS, including the creation of automated compliance-testing systems for P25 radios, coauthored the Federal Standard "Glossary of Telecommunications Terms," and is currently working on several multimedia subjective testing projects.

Arthur Webster (Webster@its.bldrdoc.gov) received the B.A. degree in English and the M.S. degree in electrical engineering. He is the chair of ITU-T SG 9, "Television and Sound Transmission and Integrated Broadband Cable Networks." For many years, he has participated in technical standards groups, most of which are devoted to the standardization of video and multimedia quality assessment methods. He was a rapporteur (in SG9 or SG12) from 1995 to 2009 and has been the cochair of the Video Quality Experts Group since its founding in 1997. For the last 21 years, he has worked for the NTIA/ITS, where he manages two technical projects. He is a Member of the IEEE and ACM and holds two U.S. patents for innovations in objective assessment of video quality.

REFERENCES

[1] ANSI-Accredited Committee T1 Contribution, "Report on an experimental combined audio/video subjective test method," *Bellcore, T1A1.5/93-104*, Red Bank, New Jersey, July 22, 1993.

[2] ANSI-Accredited Committee T1 Contribution, "Report on extension of combined audio/video quality model," *Bellcore, T1A1.5/94-141*, Red Bank, New Jersey, July 22, 1993.

[3] ANSI-Accredited Committee T1 Contribution, "Combined A/V model with multiple audio and video impairments," *Bellcore, T1A1.5/94-124*, Red Bank, New Jersey, Apr. 10, 1995.

[4] C. Jones and D. Atkinson. (1998, May 18–20). *Development of opinion-based audiovisual quality models for desktop video-teleconferencing. Proc. Rec. 6th IEEE Int. Workshop Quality of Service, Napa, CA*. [Online]. Available: www.its.bldrdoc.gov/n3/video/documents.htm

[5] "Study of the influence of experimental context on the relationship between audio, video, and audiovisual subjective qualities," *ITU-T Contribution COM12-61-E*, France Telecom/CNET, France, Sept. 1998.

[6] "Relations between audio, video, and audiovisual quality," *KPN Res.*, The Netherlands, *ITU-T Contribution COM 12-19-E*, Feb. 1998.

[7] J. G. Beerends and F. E. de Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc.*, vol. 47, no. 5, pp. 355–362, 1999.

[8] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, 2004.

[9] S.Winkler and C. Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 973–980, 2006.

[10] M. N. Garcia and A. Raake, "Impairment-factor based audio-visual quality model for IPTV," in *Proc. Int. Workshop Quality of Multimedia Experience (QoMEx)*, 2009, pp. 1–6.

[11] M. McFarland, M. Pinson, C. Ford, A. Webster, W. Ingram, S. Hanes, and K. Anderson. (2009, Sept.). *Relating audio and video quality using CIF video. NTIA TM-10-472*. [Online]. Available: http://www.its.bldrdoc.gov/pub/ntia-rpt/10-472/

[12] Advanced video coding for generic audiovisual services. *ITU-T Recommendation H.264*, Geneva, Switzerland. [Online]. Available: http://www.itu.int/en/publications/Pages/default.aspx

[13] Advanced Video Coding, ISO/IEC 14496-10–MPEG-4 Part 10, Geneva, Switzerland. [Online]. Available: http://www.iso.org/iso/store.htm

[14] M. H. Pinson, S. Wolf, and G. Cermak. (2010, Mar.). *HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss. IEEE Trans. Broadcast*. [Online]. Available: www.its.bldrdoc.gov/n3/video/

[15] Modulated noise reference unit (MNRU). *ITU-T Recommendation P.810*, Geneva, Switzerland. [Online]. Available: http://www.itu.int/en/publications/Pages/default.aspx

[16] M. H. Pinson, A. Webster, and W. Ingram. (2011, Mar.). *Preliminary investigation into the impact of audiovisual synchronization of impaired audiovisual sequences. NTIA TM-11-474*. [Online]. Available: http://www.its.bldrdoc.gov/pub/ntia-rpt/11-474/

[17] S. Jumisko-Pyykkö and T. Vainio, "Framing the context of use for mobile HCI," *Int. J. Mobile Hum. Comput. Interact. (IJMHCI)*, vol. 2, no. 4, pp. 1–28, 2010.

SP

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

[Alexander Raake, Jörgen Gustafsson, Savvas Argyropoulos, Marie-Neige Garcia, David Lindegren, Gunnar Heikkilä, Martin Pettersson, Peter List, and Bernhard Feiten]

IP-Based Mobile and Fixed Network Audiovisual Media Services

[Current approaches for quality monitoring]

This article provides a tutorial overview of current approaches for monitoring the quality perceived by users of IP-based audiovisual media services. The article addresses both mobile and fixed network services such as mobile TV or Internet Protocol TV (IPTV). It reviews the different quality models that exploit packet-header-, bit stream-, or signal-information for providing audio, video, and audiovisual quality estimates, respectively. It describes how these models can be applied for real-life monitoring, and how they can be adapted to reflect the information available at the given measurement point. An outlook gives insight into emerging trends for near- and mid-term future requirements and solutions.

Digital Object Identifier 10.1109/MSP.2011.942472
Date of publication: 1 November 2011

INTRODUCTION

IP-based video streaming or broadcast services such as video on demand (VoD), Web video (e.g., Youtube), IPTV, and mobile TV gain increasing popularity. To ensure that a service is of the desired high quality, methods are required for planning, optimizing, monitoring, and maintaining the service performance. Performance is often assessed in terms of quality of service (QoS), i.e., technical performance indicators at the protocol layer such as lost, dropped, or resent packet information and delay statistics, or at the lower layers information from streaming server logs, digital subscriber line access multiplexer

1053-5888/11/$26.00©2011IEEE

(DSLAM) information (such as link errors), radio error information (such as transmission block errors), or information from the client site based on set-top box (STB) or mobile device logs.

TO ENSURE THAT A SERVICE IS OF THE DESIRED HIGH QUALITY, METHODS ARE REQUIRED FOR PLANNING, OPTIMIZING, MONITORING, AND MAINTAINING THE SERVICE PERFORMANCE.

However, ensuring the correct functioning of the technical system is no longer sufficient, especially when the QoS performance indicators, as they are, for example, defined in the service level specifications, tolerate certain quality degradations for efficiency reasons. As a consequence, it is indispensable to assess performance in terms of user-perceived quality, often referred to as quality of experience (QoE). In that case, monitoring the QoE from the end user's perspective allows a final judgement to be made on how well the service fulfills the promises made towards the user.

Quality can be defined as the "result of [the] judgment of the perceived composition of an entity with respect to its desired composition" [1], and is internal to the user. In an ultimately valid way, it can be measured only in perception tests with human subjects. Such tests are time- and resource-consuming, and cannot be carried out to continuously monitor the quality of a running service with its numerous users. In this case, a better alternative is to use quality models that map QoS-related performance indicators, or representations of the transmitted multimedia signals, to user-perceived quality as obtained from subjects in laboratory tests.

The assessment of media-signals transmitted over telecommunication links gained broader attention when the telephone had long transformed into a large-scale service. In the 1930s and 1940s, audio quality was addressed mainly in terms of how well speech could be understood, resulting in one of the first audio-quality related models, the Articulation Index [2]. A more holistic view on speech quality was taken in the 1960s, when aspects of naturalness were addressed [3], and interactive voice communication under delay was assessed [4]. For an overview of audio and speech quality assessment see [5] and [6]. The assessment of television system quality dates back at least until the 1950s [7], with increasing research activity between the 1960s and 1980s (see [8] for an overview). Image and video quality models were first proposed in the 1970s and 1980s (see, e.g., [9]). First integration functions for combining audio and video quality estimates to an estimate of audiovisual quality were proposed in the early 1990s (see [10] for an overview).

The early assessment methods consider the degradations typical of analog systems, such as signal attenuation and noise. With the advent of digital and packet-based transmission, speech, audio, and video coding as well as error-prone transmission channels have led to new types of quality degradations, such as coding or packet loss artifacts, which require novel assessment methods. Several models have been developed in the meantime, which address the degradation of today's audiovisual media transmission systems.

Based on such models, monitoring solutions have been proposed that enable information from different locations of the end-to-end delivery chain to be converted into estimates of QoE. Such methods can be used in service probes for active or passive monitoring. These probes often form a part of a global service monitoring framework, which allows a link to be established with diagnostic QoS information. Moreover, this framework provides a service-wide picture of QoE and the sources of respective problems, for example, in terms of QoS fault isolation or performance diagnosis [11], [12].

PERCEPTUAL QUALITY TESTS

Audio, speech, video, and audiovisual quality tests should be conducted using methods recommended by the International Telecommunication Union (ITU) or European Broadcasting Union (EBU). The use of standardized methods ensures that tests are reproducible and comparable between different laboratories. Commonly used and established methods are described in P.800 [13] for speech; BS.1116 and BS.1534 [14], [15] for audio; BT.500, BT.710, P.910, and EBU-SAMVIQ for video [16]–[19]; and P.911 and P.920 for audiovisual quality tests [20], [21].

These recommendations describe different dedicated test methods, specifying aspects such as the employed rating scales and the task for which they are most appropriate (see, e.g., Table 2 in [16]). Furthermore, the standards provide guidance on the test material to be used in the test (type, duration, etc.), on the test environment and set-up, on the number of subjects to be recruited and possible methods of subject screening.

Single-stimulus methods such as the Absolute Category Rating (ACR) best reflect the everyday usage situation of watching or listening to multimedia sequences. Such methods allow a high number of conditions to be rated in a short time but are less sensitive to small quality differences. Paired or multiple comparison methods such as the Degradation Category Rating (DCR), BS.1116, Multiple Stimuli with Hidden Reference and Anchor (MUSHRA), or Subjective Assessment Methodology for Video Quality (SAMVIQ) enable more discriminative test results but can be more time-consuming than single-stimulus methods, and they do not reflect the everyday multimedia service usage either. Comparisons of methods and scales commonly used in video quality tests are reported in [22]–[24].

Hidden or direct reference stimuli are often used, typically provided by the nondegraded source sequence. The test file duration is commonly five to eight seconds for speech and ten to 16 seconds for video and audiovisual sequences, but may be several minutes when continuous evaluation methods are applied, e.g., Single-Stimulus Continuous Quality Evaluation (SSCQE) [16].

Perceptual quality tests may suffer from bias. Zielinski et al. [25] provide an overview of possible types of bias for audio listening quality tests. One type is the "response-mapping bias" that reflects how subjects map their opinion to the quality rating scale, for example, due to the quality range and degradation

types the test stimuli cover. Another source of bias can be the test interface. When conducting perceptual tests, it is vital to know these problems and to take measures to avoid them (see [25] and [26]). To the authors' knowledge, no comparable study is available for video. However, due to the fundamental perspective of [25], it can be assumed that it also applies to other modalities. An entry point into the literature on comparing video quality test methods is, for example, [24].

MODEL DEVELOPMENT USING QUALITY TEST RESULTS

Quality model development can be split into the following two parts: 1) definition of the model input and identification of respective features and 2) pooling and mapping of these features to an integral quality index. For a better separation and selection of features in terms of perceptual dimensions, it can be useful to run tests for multidimensional analysis as in [27], or qualitative tests as in [28].

In the service monitoring case, the selection of the input features depends on the available data: the signal/media, the bit stream, or both signal and bit stream (see the section "Instrumental Quality Models"). It is obvious that the best suitable input features for multimedia quality models are those that can directly be related with user perception. In case of a large number of candidate features, the finally used set may need to be identified using multivariate analysis techniques. Typical features represent (temporally and/or spatially) local artifacts or degradation due to coding or packet loss. Typical examples of pooling are: spatial pooling for image or video quality in terms of artifacts across individual frames (see e.g., [29] for image quality); temporal pooling of features or of short-term quality to an integral or longer-term quality estimate, see [6], [30], and [31] for examples for speech and video, respectively. Note that when several features associated with perceptually different types of degradations coexist, the model developer has to decide how they should best be combined. In [8] and [32], it is assumed that features that are related with a certain type of quality impairment can be mapped onto an appropriate (perceptual) quality rating scale, and that the resulting *impairment factors* I_i are additive on this scale

$$Q = \max\left(Q_{\min}\left[Q_0 - \sum_{i=1}^{n} I_i\right]\right), \tag{1}$$

with Q as the integral speech, audio, or video quality; Q_0 as the best possible base-quality in a given context; $Q_{\min}$ the lowest quality level possible on the respective model scale; and $I_i = g_i(f_i)$ the impairment factors calculated using a set of functions $g_i(\cdot)$ of the feature vectors $\boldsymbol{f}_i$, with $I_i \in [Q_{\min}, Q_0]$.

Instead of an additive formation of integral quality, other models use a multiplicative combination of the terms associated with individual impairment types, for example, in the form

$$Q_{\text{int}} = Q_0 \cdot \left(1 - \prod_i I_i\right), \tag{2}$$

where, by design, the features f_i are mapped to impairment values I_i in $[0,1]$ using a different set of functions $h_i(f_i)$.

PERFORMANCE EVALUATION OF QUALITY MODELS

Quality models aim at predictions that match perceived quality ratings as closely as possible. To ensure a guaranteed performance of a given model, two different approaches exist: Standardization of the model, as it is typically done by ITU-T or the Video Quality Experts Group (VQEG), or standardization of a procedure and testplan for model validation, as recently proposed by the IPTV Interoperability Forum (IIF) of the Alliance for Telecommunications Industry Solutions (ATIS) [33]. Model standardization within ITU or VQEG is typically conducted as a competition, where a number of proponents compete to determine the winning and then standardized model (as in case of the POLQA or PESQ development, [34], [35]), or as a partial or full collaboration. See the section "Instrumental Quality Models" for more examples of respective ITU standards. In the case of the ATIS IIF approach, not the model, but the validation procedure and criteria are standardized. Then, an independent lab conducts the respective quality tests, and evaluates the model performance. In turn, in case that the model is standardized, the test data is typically provided by the proponents or an independent lab group.

PREDICTING THE INTEGRAL QUALITY OF MULTIMEDIA SERVICES INVOLVES ESTIMATING AUDIO, VIDEO AND AUDIOVISUAL QUALITY.

In both cases, the performance of a quality model is evaluated using statistical metrics. Common metrics are the Pearson correlation coefficient (R) and the root mean square error (rmse). R is the linear correlation between the predicted and the subjective quality values on a scale from -1 to 1. The rmse measures the difference between the two data sets in terms of per-condition or per-stimulus mean square error and can be expressed on the model output or quality test scale.

In the development of ITU-T's latest full reference (FR) speech quality model POLQA (see the section "Signal-Based Models"), the modified rmse (rmse*) was used instead of rmse (see [34] for details). The performance criteria rmse and rmse* will be employed for evaluating the quality monitoring models P.NAMS and P.NBAMS currently developed by ITU-T Study Group 12. The rmse* explicitly includes the variation in the subjects' ratings

$$\text{rmse}^* = \sqrt{\frac{1}{N-d} \cdot \sum_{i=1}^{N} P_{\text{error}}(i)^2} \tag{3}$$

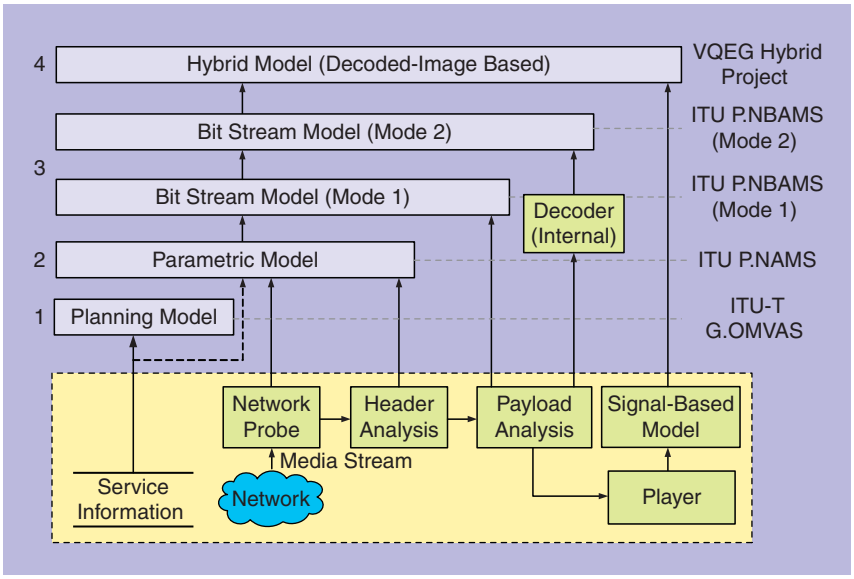
$$P_{\text{error}}(i) = \max\left(0, |\text{MOS}(i) - \widehat{\text{MOS}(i)}| - ci_{95}(i)\right). \tag{4}$$

Here, N is the number of samples; d a correction term for the case that a dth-order mapping was used to map the subjective

test data and model predictions; i is the index of the media sample; and ci_{95} the 95% confidence interval for that sample. MOS(i) and MOS(i) are the subjective and predicted quality, respectively.

INSTRUMENTAL QUALITY MODELS

Multimedia quality assessment models can be described in terms of the level at which input information is extracted. For example, information may stem from protocol headers or payload-related bit stream data. As depicted in Figure 1, models can be categorized into planning, parametric, signal-based, bit stream-based, or hybrid models, as detailed in the following sections. In addition, the models can be classified according to the amount of information they need from the original signal into no reference (NR), reduced reference (RR), and FR models. FR models have access to the original source sequence, which is compared with the processed sequence. RR models use the processed sequence together with a set of parameters extracted from the source sequence. NR models calculate quality predictions only based on the processed sequence. An overview of audiovisual quality models is given in Table 1 and in the following text. For more details on speech quality models, see [36].



[FIG1] Overview of quality assessment models that exploit different levels of information extracted from the media stream. Note: Planning and parametric models require additional service information, which they cannot extract from stream information (e.g., codec, bitrate, and loss characteristics for planning models; PLC or slicing information for parametric models; see text for details).

SIGNAL-BASED MODELS

Signal-based models exploit the decoded signal as input to calculate a quality score. Many of these models include aspects of human perception. Several ITU recommendations for signal-based models exist. In the speech quality domain, the FR-model P.862 (PESQ) [35] has been widely used over several years. The recently developed P.863 (POLQA) [34] is expected to supersede it. A respective NR speech quality model is described in [37]. For FR audio quality assessment, BS.1387 (PEAQ) [39] has been standardized.

In the video quality domain, J.144 [40] specifies models for FR SDTV quality assessment, J.341 [42] for FR HDTV

[TABLE 1] OVERVIEW QUALITY MODELS, SEE TEXT FOR DETAILS.

MODALITY	FR	RR	NR	P	BS(1)	BS(2)	IMPAIRMENT	OUTPUT	EXAMPLES
S	X						C, L	QUAL.	P.862/PESQ [35], P.863/POLQA [34]
S			X				C, L	QUAL.	P.563 [37]
S			X	X			C, L	QUAL.	G.107/E-MODEL [32], P.564 [38]
A			X	X			C, L	QUAL.	G.OMVAS, P.NAMS
A	X						C	QUAL.	BS.1387/PEAQ [39]
V	X						C	QUAL.	J.144 [40]
V	X						C, L	QUAL.	J.247 [41], J.341 [42]
V		X					C, L	QUAL.	J.246 [43]
V			X	X			C, L	QUAL.	P.NAMS; [44]–[49]
V			X	X			C, L, R	QUAL.	P.NAMS; [50]
V			X		X	X	C, L, R	QUAL.	P.NBAMS (MODE 1, MODE 2); R FOR MOBILE CASE
V			X		X		C	QUAL.	[51]–[53]
V			X			X	C	QUAL.	[54]
V	X[1]	X[1]	X[1]		X		L	VIS.	[55], [56]
V			X			X	L	MSE	[57]
A/V			X	X			C, L	QUAL.	G.1070 [58]; P.NAMS; [10]; G.OMVAS

S: Speech; A: Audio; A/V: Audiovisual; V: Video; P: Parametric/planning; BS(1): Bit stream mode 1, ITU-T P.NBAMS activity; BS(2): P.NBAMS Bit stream mode 2. Impairment: [C: Coding; L: Loss; R: Rebuffering]; Output: [Qual.: Quality; Vis.: Visibility; MSE: Mean Squared Error]. X1:Respective models available in NR, RR, and FR versions.

quality assessment, and J.246 and J.247 [41], [43] for RR and FR video quality assessment, respectively. See [9] for an overview of additional signal-based models and measures such as peak signal to noise ratio (PSNR). So far no audio-visual signal-based model has been standardized. Although the accuracy of the quality estimation for the video models is high, their high processing demands make them less well suited for larger-scale live network monitoring. Instead, they are better suited for off-line assessment or active monitoring in specific end-point test equipment. RR and FR models, which have the best accuracy of the signal-based models, must also deal with the nontrivial problem of synchronizing the reference signal to the decoded signal (see, e.g., [59] for a comparison of some video-related methods).

THE VISIBILITY AND THUS VIDEO QUALITY IMPACT OF PACKET LOSS HIGHLY DEPENDS ON DECODER PLC, AND THE SPATIOTEMPORAL COMPLEXITY OF THE CONTENT.

PARAMETRIC AND PLANNING MODELS

Parametric quality models predict the impact of coding and IP network impairments on multimedia QoE. Models of this type do not access the packet payload; instead, they use information extracted from packet headers, i.e., transport stream (MPEG-2 TS), Real-Time Transport Protocol (RTP), or packetized elementary stream (PES) headers, depending on the level of encryption. Parametric models do not explicitly address aspects such as the source quality or encoder or player implementations. To include these, service information (Figure 1) can be used to chose from different sets of curve-fitting coefficients. Parametric audio, video, and audiovisual quality models are currently developed and standardized in ITU-T Study Group 12 under the provisional name P.NAMS. Two different P.NAMS models are developed; one for the low bit-rate application area including mobile TV, and one for the high bit-rate application area that includes services such as IPTV.

Planning models can be considered as a variant of parametric models, where the input information is not acquired from an existing service, but is estimated based on service information available during the planning phase. The ITU-standardized E-model [32] is a prominent example for speech services. Another example is ITU-T Rec. G.1070 [58], a planning model for interactive video-telephony services. A corresponding model for streaming media is currently developed by ITU-T under the provisional name G. OMVAS.

Parametric models that are applied for service monitoring must provide media sequence-related quality estimates. In both mobile and fixed services, the quality degradations captured during monitoring relate to compression, packet loss, and delay, with delay leading to dropped packets or client rebuffering. How these degradations are handled by parametric models is discussed next.

COMPRESSION

Degradation of audio or video quality due to compression is mainly determined by the number of bits allocated to different audio or video frames, respectively. For video, frame rate and resolution are additional factors that impact quality. In case of video compression, the bit allocation mainly depends on

- the video resolution and frame rate
- the employed codec type and profile
- the codec implementation
- the targeted bit-rate
- the group of picture (GOP) structure
- the spatiotemporal complexity of the transmitted video content.

In case of audio, the bit allocation depends on the spectral bandwidth and the dynamics of the signal. Speech codecs such as Adaptive Multirate (AMR) employ a speech source model as the basis for quantization, whereas audio codecs such as Advanced Audio Coding (AAC) and MPEG-1 Layer 3 (MP3) exploit auditory masking.

The media resolution, sample rate, and employed codec type and profile are usually known in terms of service information and do not need to be extracted from packet headers (see Figure 1). At transport level, the spatiotemporal complexity of the video content is unknown. However, since high spatial complexity results in larger I-frames, while high temporal complexity usually yields larger P- and B-frames, a prediction of video complexity from frame size statistics was proposed [60].

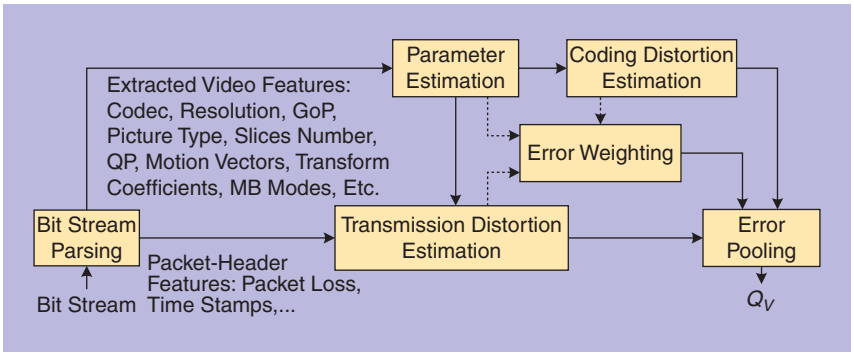
Most parametric video quality models estimate the impact of video compression using only, for a given resolution, the bit-rate, codec type, and profile, and possibly the GOP structure [44]–[46]. In the case of mobile services, the video frame-rate is used in addition [58] (wireline services such as IPTV have a fixed frame-rate). Audio quality models typically use the codec type and bitrate as input. The quality is then basically modeled as an exponential, logarithmic, or power function of a combination of the respective input parameters.

PACKET LOSS

The packet loss rate is commonly used in the literature as model input for describing the amount of lost packets [44], [46]. Bursty losses are described with the average number of packets lost in a row, the burst density (the fraction of packets lost in a burst) and the burst duration [47], [48], or the packet loss frequency [45]. For audio, consecutive losses can affect the efficiency of the packet loss concealment (PLC) method, which can typically handle one or two subsequently lost audio frames.

The quality impact of packet loss highly depends on the PLC method employed by the decoder. Video PLC can be classified as slicing or freezing. A slice typically corresponds to a certain area of the picture that is encoded independently of

other areas in that picture. If affected by loss, the decoder conceals the affected slice area with data from adjacent regions or frames, or motion compensated content. In case of freezing concealment, the picture freezes until the next intact reference frame. Simple frame repetition may be used for audio PLC. More advanced methods make predictions based on surrounding audio frames, and typically introduce less frame-boundary-related artifacts.



[FIG2] **Architecture of a bit stream-based video quality assessment model.**

In general, increasing the number of slices increases the robustness of the video sequence to packet losses (since it reduces the spatial extent of the loss), but reduces the coding efficiency, due to added slice headers and interrupted predictive coding at slice boundaries. Information such as the PLC, or the number of slices per frame in case of transport streams cannot be extracted from packet headers, and thus has to be available to the parametric model as side information. In case of slicing, the spatiotemporal extent of the loss depends on the frame type, slice size and number of packets per frame. The loss-related degradation propagates until the next intact reference frame that does not reference other erroneous frames. If the frame type can be identified, the loss-impact can be estimated using parameters such as the number of distorted (invalid) frames, see, e.g., [49].

REBUFFERING

Rebuffering for video and audio is an important degradation caused by packet loss or a too-narrow streaming bandwidth. This parameter is common in YouTube-type streamed audiovisual media and Web radio or Spotify-like audio. The decoder decides whether to use rebuffering or freezing/slicing when a late loss occurs. YouTube will rebuffer until the data is received to continue playback. Hence, if one packet is lost or simply late, the client will ask for a retransmission until a sufficient portion of the sequence is received to continue playback. A parametric video quality model comprising rebuffering is described in [50]. If a precise measurement is desired, rebuffering information must be measured in the client and reported back to the monitoring model location. If this is not achievable, it is possible to estimate buffer behavior depending on the acquired packets and the encoded media bitrate, providing a rough estimate of the rebuffering degradation.

VIDEO BIT STREAM MODELS

In addition to transport layer information, bit stream quality models exploit information from the elementary stream. Video bit stream quality models for IPTV and mobile applications are currently being standardized by ITU-T in the Study Group 12 workitemP. NBAMS. Twodifferentmodesareunderconsideration:in Mode 1, the model does not fully decode the payload, but only parses the bit stream and extracts features; in Mode 2, the model can fully decode the bit stream (including the inverse transformation), and use the additional information from the reconstructed pixel values.

The architecture of a bit stream quality assessment model is depicted in Figure 2. Initially, the bit stream headers are parsed to extract transport-related information such as TS and/or RTP time stamps and sequence numbers for packet loss detection. Additionally, the payload of the video bit stream is parsed, and different features are extracted. The most important features are the picture type, the number of slices, the quantization parameter (QP), the motion vector and type of each macroblock (MB) and its partitions, and the transform coefficients of the prediction residual. Subsequently, these features are further processed to compute the inputs to the coding distortion and transmission distortion estimation modules, and to determine the parameters employed in the perceptual error weighting module, which can include aspects of the human visual system, such as visual attention. Typically, a bit stream model separately evaluates the degradation due to compression or preprocessing of the sequences and the distortions due to network impairments. These distortions can be locally evaluated (e.g., at the frame or GOP level) and are pooled to provide a unique estimation of the quality of the video sequence (Q_V in Figure 2).

In the following, a number of example model algorithms are summarized; some general considerations on quality monitoring including content information can be found in [61]. A coding error estimation method based on the distribution of transform coefficients was proposed in [52]. In this approach, the perceptual model is based on the spatiotemporal contrast sensitivity function (CSF) applied to the transform domain for local error weighting. Furthermore, the spatial information (SI) and temporal information (TI) parameters defined in P.910 [18] are explicitly employed for the coding distortion assessment in [54]. Another method based on the extraction of spatial and temporal features from the bit stream and the prediction of video quality based on blocking and flickering artifacts was presented in [51]. Content-dependent parameters were selected from a set of spatiotemporal features derived from the video bit stream, such as the AC transform coefficients, QP, and the motion vectors in the method presented in [53].

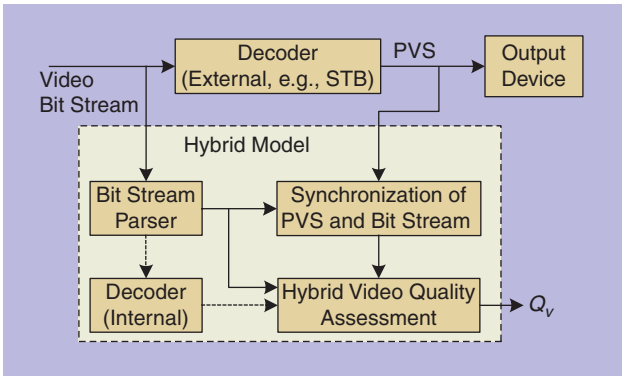
For the evaluation of quality degradations due to network impairments, the models should take into account that packet losses are not equally important and produce a different distortion depending on the pattern of loss events, their position within the sequence, the video content, and neighboring information that may mask the induced error. The visibility of packet loss in H.264/AVC streams was investigated in [55]. Based on different features extracted from the video bit stream, a general linear model (GLM) was proposed to predict the visibility of packet loss degradation.

The concept of mean time between failures was introduced in [62] to evaluate video quality based on failure statistics. The automatic video quality (AVQ) method evaluates compression and transmission artifacts, using the quantization step size and activity in scenes, and the perceptibility of video artifacts [56]. Finally, a no-reference method for quality monitoring based on the estimation of the mean squared error distortion at the MB level was presented in [57].

In [63], a model that directly incorporates motion information in an information theoretic framework was presented. Furthermore, visual attention is modeled in [64] from motion and contrast information for error-weighting based on saliency maps and pooling of global and local quality.

HYBRID VIDEO QUALITY ASSESSMENT MODELS

The block diagram of a hybrid video quality assessment model is depicted in Figure 3. Similar to the bit stream models, which may perform full decoding of the received bit stream, the hybrid video quality assessment models exploit information from the packet headers, the elementary stream, and the reconstructed pictures. The critical difference is that the information for the reconstructed pictures is obtained from the processed video sequence (PVS) generated by an external decoder (e.g., an STB in case of IPTV) and not from an internal decoder within the model. Thus, the reconstructed pictures used in the model are identical to the ones observed by the viewers, and there is no mismatch between the assumed and the actual packet loss handling and concealment method. Nevertheless, the model may also employ an internal decoder (like in Mode 2 of video bit stream models), since certain useful decoder-level information is not available from closed systems such as most STBs. If the bit stream features are not time-aligned with the picture-related features, the quality predictions become inaccurate. Thus, a crucial step in the hybrid video quality model is the time-alignment of the video bit stream with the decoded reconstructed images. Hybrid quality assessment models are currently examined in the Hybrid Perceptual/Bit stream project of VQEG [65]. In addition, the Joint Effort Group (JEG) is a new activity of VQEG that proposes an alternative, collaborative action for NR hybrid video models [66].



[FIG3] **Block diagram of a hybrid video quality assessment model.**

3-D VIDEO QUALITY MODELS

Models of three-dimensional (3-D) video quality must—in addition to the quality attributes for two-dimensional (2-D) video—consider 3-D-related quality aspects including depth perception, the naturalness of a scene, and the visual discomfort. Additional artifacts may also be present such as geometric distortions due to camera limitations [67], color differences, and spatial and temporal offsets between the left and the right view, for example, due to the employed (2-D-based) loss concealment. Different transmission techniques are currently being discussed for 3DTV applications, such as (the stereoscopic) frame-packing, which typically means side-by-side inclusion of two views in one frame and subsequent 2-D encoding, video plus depth encoding, or multiview coding (MVC) [68]. In case of stereoscopic coding approaches, instrumental assessment methods as for 2-D can be adopted [69]; the quality of each view can be calculated with a 2-D video model, and an additional modeling approach is used for the depth-related coding and transmission distortions. In case of video plus depth encoding, the specific reconstruction distortions have a very strong impact on quality and must be addressed with completely new types of models.

AUDIOVISUAL QUALITY MODELS

Audiovisual quality depends on the audio quality, the video quality, their interaction, and audiovisual impairments such as lip-sync problems. The simplest solution for estimating audiovisual quality is to use a function of audio and video quality estimates, and calculate the audiovisual quality score regardless of what type of degradation affects the audio and video scores. Linear regression with or without interaction between audio quality Q_A and video quality Q_V were proposed as quality integration functions (e.g., [10] and [70]), with respectively chosen coefficients α_i to calculate audiovisual quality Q_{AV}

$$Q_{AV} = \alpha_1 + \alpha_2 \cdot Q_A + \alpha_3 \cdot Q_V + \alpha_4 \cdot Q_A \cdot Q_V. \quad (5)$$

More detailed predictions can be achieved by using intermediate audio and video features that underlie the overall audio and video scores, and map these to an audiovisual quality score. The increased accuracy, however, comes with an increased complexity of the model. For a comparison of approaches, see [10].

MONITORING WIRELINE AND MOBILE SERVICES

The following description focuses on the application of quality models for monitoring QoE of IP-based audiovisual services. Respective tools consider the impact of IP network impairments on the quality of mobile audiovisual streaming and IPTV applications over transport formats such as User Datagram Protocol (UDP), RTP, Transmission Control Protocol (TCP), and Hypertext Transfer Protocol (HTTP), employing further protocols and data formats such as MPEG-2 TS, and various Internet Engineering Task Force (IETF) payload formats.

Three main criteria can serve for classifying and selecting the best suitable approach in a given context:

- *Target application area*
  - The service to be assessed: on demand streaming, broadcast, interactive, real-time communication etc.
  - The implementation of the service in terms of the protocols, audio and video resolutions, codecs, the manufacturer(s), and respective (proprietary and nonproprietary) implementations
- *Implementation of the assessment approach*
  - The locations of the quality model and of the probe for input data acquisition along the distribution chain
  - The monitoring may be done during service operation in a nonintrusive, i.e., passive manner, or off-line, as active (intrusive) monitoring, with dedicated equipment that inserts test traffic
- *Employed quality model*
  - The modality the model applies to: audio, video, or audiovisual
  - The type of information available to the model, for example, packet headers, encoded bit stream information, decoded signals, and the amount of reference information used (NR, RR, and FR models).

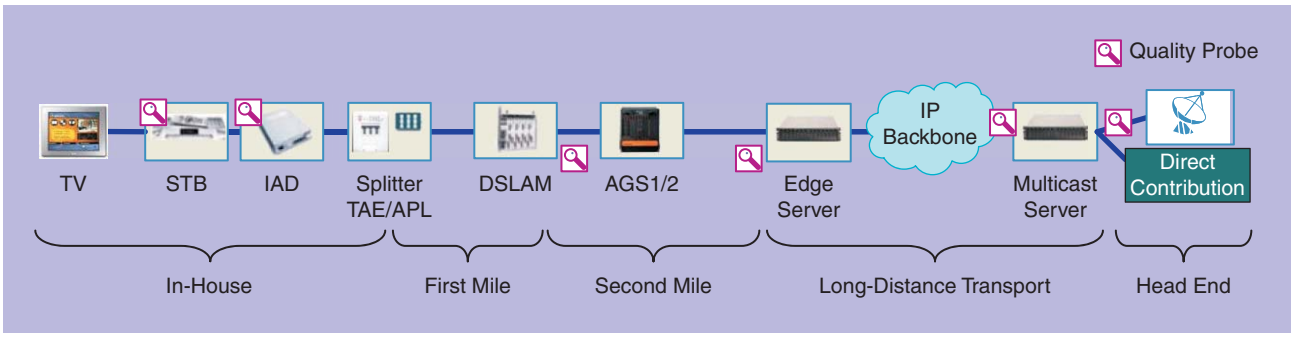
The quality model may be deployed directly in the measurement probe that acquires the model input information, or independently in a different location. Possible locations are the end-user devices or other midnetwork monitoring points. The locations of the model and measurement probe determine the mode of operation. Four such modes of operation are considered by ITU-T Study Group 12 in the P. NAMS development. They are referred to by a two-letter code, each one respectively for the probe and the model location, with three possible location designation letter choices: N for network (head-end, server, or another point along the delivery chain), C for client, and B for both. The four modes of operation are:

- NN Probe and model are located in the network. The model can use known service information (see Figure 1) and streaming data captured at this point.
- BN Probe located in the network and client—model located in the network. In this case, the information about the end point needs to be collected through measurement reporting protocols such as [71] and [72].
- CN Probe is located in the client—model is located in the network. All model input information other than service information known in prior is measured at the client and transmitted to the model through measurement reporting protocols.
- CC Probe and model are located in the client.

Another important issue for service deployment is how to increase error resilience. Common methods are forward error correction (FEC) with redundant information for error correction added to the stream, and automatic repeat request (ARQ), where missing information is requested and resent from a server. Depending on where in the delivery chain the error correction method is deployed, and where the monitoring information is captured, error correction has to explicitly be taken into account for QoE predictions. Two principle approaches are conceivable to capture this case:

- The packet stream is converted into a corrected stream that reflects an assumed behavior of FEC or ARQ.
- The input parameters/features to the model are converted into corrected values that simulate the effect of prior application of FEC or ARQ.

For different services and network structures, an adequate selection of measurement points is required (see Figure 4). The available information may vary, and not all model types can be applied in all situations. For example, if selective decryption is not efficiently feasible, the quality of encrypted media streams can only be assessed with parametric models using header information.



[FIG4] IPTV end-to-end delivery chain. Potential monitoring points are indicated by the "quality probe" symbols.

FIXED-NETWORK MEDIA SERVICE MONITORING

A typical wireline IPTV end-to-end service architecture for linear TV is depicted in Figure 4. It consists of a head end; a central play-out service (here based on a multicast server); an IP backbone for long-distance transport; some additional distributed servers at the network edges; the second mile aggregation network and switches (AGS); the first mile access network with the DSLAM; the Internet access devices (IADs); and the in-house network with the user's end device.

AN ESSENTIAL ENTITY OF THE QUALITY MONITORING SYSTEM IS A CENTRAL SERVICE MANAGEMENT THAT STEERS THE DISTRIBUTED PROBES AND COLLECTS AND DIGESTS THE MEASUREMENT RESULTS.

A complete monitoring system typically provides various types of information on system performance. The following description concentrates on monitoring the IPTV streams in terms of audiovisual quality. A quality monitoring probe can be applied at different stages of the delivery chain; see Figure 4. In the head end, an FR signal-based or hybrid model can provide an optimal basis for comparing measurement results with that of following measurement points. The next potential measurement points are located after the play-out server, at the different edge servers, at the AGS, and at the DSLAM. Here, parametric or bit stream quality models could be applied. Probes can also be placed in the homes, inside the IAD, in the STBs, or as an additional box linked between the IAD and the STB, monitoring all TV streams requested from the STB. The best placement of the probes depends on the implementation details of the whole system.

An essential entity in the architecture of the quality monitoring system is a central service management that steers the distributed probes and collects and digests the measurement results. The IPTV-service example is based on multicast transmission of RTP packets carrying MPEG-2 transport streams. For optimizing the channel switching time, parts of the stream are sent as unicast traffic coming from the edge server. Lost packets are retransmitted via unicast as well (ARQ), and need to be considered during QoE monitoring. A further challenge is that the payload of the MPEG-2 transport stream is encrypted for many of the streams. The quality monitor probe has to support the dissection of the whole IP traffic and extract the IPTV-related packets from it. The different IPTV streams are then analyzed in depth. Depending on the mode of operation, parametric, bit stream, or hybrid models according to the section "Instrumental Quality Models" are employed at one or several locations.

The quality model needs to be tailored to the processing power of the device it is implemented in. A parametric model can run on the IAD, while on the STB, a bit stream or even hybrid model could be implemented (CC mode). Alternatively, a lightweight probe on the client can send information to a more complex model located in the network (BN, CN modes). For a broad deployment, a solution based on standards is required. The upcoming P.NAMS and P.NBAMS standards together with appropriate standards for the communication between the end-user devices and the service monitoring systems, such as TR-069 and TR-135 [71], [73], optimally support the integration of QoE monitoring solutions in the IAD and STB. In this way, the STB or the IAD can measure what the end user actually experiences, and these measurements can be collected by the service management system in a cost-efficient manner. As an alternative to a probe or model on the STB or IAD, an additional box can be used in between these devices that sends quality reports to the service management system. Because of the extra costs, such end-user probes are only placed in exemplary users' homes, or in cases where a service engineer has to solve specific problems.

MOBILE MEDIA SERVICE MONITORING

For mobile media, the same type of monitoring as in the fixed case is principally possible, with quality probes located at different nodes in the distribution network. However, most of the quality impact is usually related to problems in the last wireless hop, which makes network-internal measurements less useful for assessing user-perceived quality. Hence, the only way to get a complete picture of QoE is to make the measurements directly in the mobile client, and report them back to the system. For 3GPP devices, QoE reporting was standardized in Release 6 for RTP-based streaming, and extended in Release 8 and 9 to HTTP streaming, progressive download and multimedia telephony (MMtel) [74], [75].

During a multimedia session, the mobile client continuously collects quality-related measurements, and periodically reports these back to a QoE reporting server that is run by the operator or the service provider. By using the reported QoE metrics as input to a parametric multimedia model, a good estimate of the end-user quality can be obtained. The data collected by the QoE reporting server is typically combined with other network or node measurements so that the cause of any quality degradation can be identified. Note that each mobile client reports QoE metrics regarding the received media, so to get the full picture, the QoE server can combine the reports from different clients.

Practical implementations typically address RTP-based streaming, HTTP streaming, and MMtel. The mobile client can be preconfigured via Open Mobile Alliance Device Management (OMA-DM) [76] with a default QoE reporting configuration. This default configuration is used by the mobile client if no session-related QoE configuration is specified. Session-related QoE reporting could, for instance, be based on a QoE configuration in the session initiation sequence provided by the RTP streaming server. The default configuration can specify that QoE reports shall be sent for selected sessions, all sessions, or for a random subset (for example, 5%). It is possible to specify which metrics are reported, and how often these are measured. In this way, the QoE reporting can be fine-tuned to achieve a good understanding of service QoE, without consuming unnecessary uplink radio capacity.

FUTURE TRENDS AND CHALLENGES

The standardization of parametric and bit stream-based as well as hybrid models will allow current monitoring requirements to

be complemented by standardized models. Respective ITU-T standards are expected to be available by mid-2012.

In the meantime, the media landscape is changing and expanding very rapidly, with more complex and rich service offerings as well as converged mobile-fixed solutions brought to the general public. People will expect the same or even better media quality as with the legacy solutions, and efficient and accurate methods for assessment of end-user service quality is one key element in this scenario.

MODEL DEVELOPMENT

One dimension of evolution is seen in the model development as such. The first aspect along these lines considers the ecological validity of the subjective test data that underlie all current models: to guarantee controlled laboratory settings, ecologically valid scenarios are not taken into account when asking subjects for quality ratings. However, the design and context of most tests and thus the respective model predictions differ considerably from how multimedia services are used in practice. Short sequences ranging from ten to 30 s duration are typically used instead of more TV- or video-typical viewing durations; the lab-test environment is physically treated to yield controlled optical and acoustical conditions, and thus lacks the naturalness of home or outdoor media consumption; the modeling data consists of conscious quality ratings rather than more hedonic user reactions, and so on. These drawbacks are deliberately chosen for the benefit of highly controlled settings and thus reproducible results. However, extrapolation or mapping to the actual usage scenario will help service providers and manufacturers to have a clearer picture of the relevance and value of certain model predictions. In [77], a first attempt was made along these lines.

Alternative methods are especially required for 3-D media quality assessment. While current test methods and models mainly provide fidelity-type data, more hedonic features or aspects of the "fidelity of the experience" such as media immersion need to be addressed. For 3-D, this is of cardinal relevance: for example, for video at a fixed bitrate, a viewer in a lab-test is likely to prefer a 2-D version over a 3-D version in terms of sheer image quality, but at home may still chose the 3-D video due to its higher hedonic appeal.

Accordingly, also for quality modeling, 3-D audio and video will be subjects of further research and development: although first considerations exist on how current 2-D video quality models may be adapted to 3-D, real 3-D models that are at par with their 2-D counterparts in terms of prediction accuracy are not yet available. This is partly due to the specific implications of 3-D video and related problems such as visual fatigue and diplopia [78], but also due to the employed coding and transmission strategies; see the section "3-D Video Quality Models." The situation for 3-D audio quality models is a very similar one. Subjective tests have indicated that timbral features are more important than spatial ones [79]. First respective models have recently been developed, but are signal based and cannot be used in a monitoring context [80]. For the ultimate goal of 3-D audiovisual quality prediction, no models are currently available.

ALTERNATIVE METHODS ARE ESPECIALLY REQUIRED FOR 3-D MEDIA QUALITY ASSESSMENT.

IMPLEMENTATION OF SERVICE QUALITY MEASUREMENTS

Ideally, all new services or new types of devices should include built-in support for quality assessment and respective reporting. This will enable network operators and service providers to use such quality measurements as one important input to the optimization and trouble-shooting of their services and content-delivery networks. This requires standardization of not only media quality models, but also of concepts and protocols, which together can be used to efficiently implement quality assessment and client quality feedback reporting.

Existing approaches for integrating QoE and QoS data from different sources need to be extended [11], and the advantages of data warehouse-based solutions in combination with semiautomatic or automatic optimization of networks need to be exploited in more depth. This way, existing methods for fault localization, control, and network tomography will be complemented by respective QoE data, enabling a more comprehensive and QoE-based service monitoring, possibly considering aspects of user behavior and preferences. Such a multiservice perspective that includes user-related information will enable service-centric optimization and control of networks, more dedicated resource and service management, and to efficiently plan and implement comparable new services.

AUTHORS

Alexander Raake (alexander.raake@telekom.de) received his doctoral degree in electrical engineering and information technology from Ruhr-University Bochum, Germany, in 2005, and his electrical engineering diploma from RWTH Aachen, Germany, in 1997. From 1998 to 1999 he was a researcher at EPFL, Switzerland. Between 2004 and 2009 he held postdoc and senior scientist positions at LIMSI-CNRS, France, and Deutsche Telekom Laboratories, Germany, respectively. Since 2009, he has been an assistant professor at Deutsche Telekom Laboratories, Technical University Berlin. His research interests are in multimedia technology and QoE. He has been active in ITU-T since 1999 and is currently corapporteur of Q.14/12 on audiovisual quality.

Jörgen Gustafsson (jorgen.gustafsson@ericsson.com) obtained his M.Sc. degree from Linköping University, Sweden, in 1993. He then joined Ericsson Research, Luleå, Sweden, where he currently works as a master researcher in the field of network management. His research interests are in network management, performance monitoring, QoE, and end-user quality perception. Since 2001, he has been involved in ITU-T standardization, where he currently acts as corapporteur for question Q.14/12 on monitoring models for audiovisual services.

Savvas Argyropoulos (savvas.argyropoulos@telekom.de) received the diploma and Ph.D. degrees in electrical and

computer engineering from the Aristotle University of Thessaloniki, Greece, in 2004 and 2008, respectively. He is currently a research scientist at Deutsche Telekom Labs, Technical University of Berlin, Germany. He was also a research associate at the Informatics and Telematics Institute, Centre for Research and Technology Hellas. His research interests include video quality assessment, video coding and transmission, and distributed source coding. He received the 2008 European Biometrics Research Award. He is a Member of the IEEE.

Marie-Neige Garcia (marie-neige.garcia@telekom.de) holds a master's degree in engineering from the Electronics Superior Institute of Paris (France). She is currently pursuing her Ph.D. degree on audio-visual quality modeling of IPTV services at Deutsche Telekom Labs, Technical University of Berlin, Germany. From 2004 to 2006, she was in charge of the evaluation of text-to-speech systems and video technologies in both French and European projects at the Evaluation and Language Resources Distribution Agency, France. Her research topics are video and audiovisual quality assessment and quality modeling. She contributes to the ITU-T SG12 Q13 and Q14 activities and to VQEG.

David Lindegren (david.lindegren@ericsson.com) is an experienced researcher in the department of Wireless Access Networks at Ericsson Research in Luleå, Sweden. He joined Ericsson immediately after receiving his M.Sc. degree from the Luleå University of Technology in 2007. His main research interests are end-user subjective quality assessment for multimedia services together with radio measurements. He has been involved in related ITU-T standardization since 2008.

Gunnar Heikkilä (gunnar.heikkila@ericsson.com) is a senior specialist in the Department of Wireless Access Networks at Ericsson Research, Luleå, Sweden. He joined Ericsson in 1987 after obtaining an M.Sc. degree at Luleå Technical University. His current research areas are service quality measurements and end-user quality perception. He is also active in related standardization, mainly within 3GPP and ETSI.

Martin Pettersson (martin.m.pettersson@ericsson.com) is working as a senior researcher in the multimedia department at Ericsson Research in Stockholm, Sweden. He started at Ericsson after obtaining his M.Sc. degree in information and communications technology at Lund University, Sweden, in 2005. His main research interests are in subjective and objective video quality assessment and 3-D video. He is also interested in research of image and video compression. He has been involved in ITU-T standardization and VQEG activities since 2008.

Peter List (peter.list@telekom.de) graduated in applied physics in 1985 and received the Ph.D. degree in 1989 from the University of Frankfurt/Main, Germany. Since 1990, he has been with Deutsche Telekom, currently as a senior expert for video coding and quality at Deutsche Telekom Laboratories. He has actively contributed to international standardization of video compression technologies in MPEG, ITU, and several European projects.

Bernhard Feiten (bernhard.feiten@telekom.de) received the diploma and Dr.-Ing. degrees in electronic engineering from the Technische Universität Berlin in the field of speech synthesis, psychoacoustics, and audio bitrate reduction. He worked as an assistant professor at Technische Universität Berlin in the field of communication science and digital signal processing. Since 1996, he has been with Deutsche Telekom, currently as senior expert and project leader at Deutsche Telekom Laboratories. His research fields comprise audio and video coding algorithms, broadcasting applications, high-quality Internet media distribution and streaming, and QoE monitoring. He is actively working in standardization, mainly 3GPP SA4.

REFERENCES

[1] U. Jekosch, *Voice and Speech Quality Perception—Assessment and Evaluation*. Berlin: Springer-Verlag, 2005.

[2] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, no. 1, pp. 90–119, 1947.

[3] R. Bücklein, "Hörbarkeit von Unregelmäßigkeiten in Frequenzgängen bei akustischer Übertragung," *Frequenz*, vol. 16/1962, pp/ 103–108, 1962.

[4] R. M. Krauss and P. D. Bricker, "Effects of transmission delay and access delay on the efficiency of verbal communication," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 286–292, 1966.

[5] S. Bech and N. Zacharov, *Perceptual Audio Evaluation*. Hoboken, NJ: Wiley, 2006.

[6] A. Raake, *Speech Quality of VoIP—Assessment and Prediction*. Hoboken, NJ: Wiley, 2006.

[7] P. Mertz, A. Folwer, and H. Christopher, "Quality rating of television images," *Proc. Inst. Radio Eng.*, vol. 38, pp. 1269–1283, 1950.

[8] J. Allnatt, *Transmitted-Picture Assessment*. New York: Wiley, 1983.

[9] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, 2008.

[10] M.-N. Garcia, R. Schleicher, and A. Raake, "Impairment-factor-based audio-visual quality model for IPTV: Influence of video resolution, degradation type, and content type," *EURASIP J. Image Video Processing*, vol. 11, no. 629284, pp/ 1–14, 2011.

[11] A. Mahimkar, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and Q. Zhao, "Towards automated performance diagnosis in a large IPTV network," in *Proc. ACM SIGCOMM*, 2009, pp. 1–12.

[12] A. Begen, C. Perkins, and J. Ott, "On the use of RTP for monitoring and fault isolation in IPTV," *IEEE Network Mag.*, vol. 24, no. 2, pp. 14–19, 2010.

[13] ITU, "Methods for subjective determination of transmission quality," ITU-T Rec. P.800, Int. Telecomm. Union, Geneva, Switzerland, 1996.

[14] ITU, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," ITU-R Rec. BS.1116-1, Int. Telecomm. Union, Geneva, Switzerland, 1997.

[15] ITU, "Method for the subjective assessment of intermediate quality levels of coding systems (MUSHRA)," ITU-R Rec. BS.1534-1, Int. Telecomm. Union, Geneva, Switzerland, 2003.

[16] ITU, "Methodology for the subjective assessment of the quality of television pictures," ITU-R Rec. BT.500-12, Int. Telecomm. Union, Geneva, Switzerland, 2009.

[17] ITU, "Subjective assessment methods for image quality in high-definition television," ITU-R Rec. BT.710-4, Int. Telecomm. Union, Geneva, Switzerland, 1998.

[18] ITU, "Subjective video quality assessment methods for multimedia applications," ITU-T Rec. P.910, Int. Telecomm. Union, Geneva, Switzerland, 1999.

[19] ITU, "SAMVIQ-Subjective assessment methodology for video quality," EBU-SAMVIQ, Int. Telecomm. Union, Geneva, Switzerland, 2003.

[20] ITU, "Subjective audiovisual quality assessment methods for multimedia applications," ITU-T Rec. P.911, Int. Telecomm. Union, Geneva, Switzerland, 1998.

[21] ITU, "Interactive test methods for audiovisual communications," ITU-T Rec. P.920, Int. Telecomm. Union, Geneva, Switzerland, 2000.

[22] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal Process.*, vol. 77, no. 1, pp. 1–9, 1999.

[23] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE*, vol. 5150, p. 573, 2003.

[24] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. J. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, 2011.

[25] S. Zieliński, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests—A review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.

[26] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Norwell, MA: Kluwer, 2000.

[27] M. Waltermann, A. Raake, and S. Möller, "Quality dimensions of narrowband and wideband speech transmission," in *Acta Acustica utd with Acustica*, vol. 96, no. 6, pp. 1090–1103, 2010.

[28] S. Jumisko-Pyykko, J. Hakkinen, and G. Nyman, "Experienced quality factors—Qualitative evaluation approach to audiovisual quality," in *Proc. SPIE*, vol. 6507, 2007, p. 6507M.

[29] Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE ICIP*, 2006, pp. 2945–2948.

[30] B. Weiss, S. Möller, A. Raake, J. Berger, and R. Ullmann, "Modeling call quality for time-varying transmission characteristics using simulated conversational structures," *Acta Acustic utd with Acustica,* vol. 95, no. 6, pp. 1140–1151, 2009.

[31] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Select. Topics Signal Proc.*, vol. 3, no. 2, pp. 253-265, 2009.

[32] ITU, "The E-model, a computational model for use in transmission planning," *ITU–T Rec. G.107, Int. Telecomm. Union*, Geneva, Switzerland, 2009.

[33] ATIS-IFF, "Technical report on a validation process for IPTV perceptual quality measurements," *ATIS-0800035, Alliance for Telecommunications Industry Solutions, IPTV Interoperability Forum,* 2010.

[34] ITU, "Perceptual objective listening quality assessment (POLQA)," *ITU–T Rec. P.863, Int. Telecomm. Union*, Geneva, Switzerland, 2010.

[35] ITU, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU–T Rec. P.862, Int. Telecomm. Union*, Geneva, Switzerland, 2001.

[36] S. Möller, W.-Y. Chan, N. Côté, T. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Mag.*, vol. 28, no. 6, pp. 18–28, 2011.

[37] ITU, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," *ITU–T Rec. P.563, Int. Telecomm. Union*, Geneva, Switzerland, 2004.

[38] ITU, "Conformance testing for narrowband voice over IP transmission quality assessment models," *Int. Telecommun. Union, CH-Geneva, ITU–T Rec. P.564*, 2007.

[39] ITU, "Perceptual evaluation of audio quality (PEAQ)," *ITU–R Rec. BS.1387, Int. Telecomm. Union*, Geneva, Switzerland, 2001.

[40] ITU, "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference," *ITU–T Rec. J.144, Int. Telecomm. Union*, Geneva, Switzerland, 2004.

[41] ITU, "Objective perceptual multimedia video quality measurement in the presence of a full reference," *ITU–T Rec. J.247, Int. Telecomm. Union*, Geneva, Switzerland, 2008.

[42] ITU, "Objective perceptual multimedia measurement of HDTV for digital cable television in the presence of a full reference," *ITU-T J.341, Int. Telecomm. Union*, Geneva, Switzerland, 2010.

[43] ITU, "Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference," *ITU–T Rec. J.246, Int. Telecomm. Union*, Geneva, Switzerland, 2008.

[44] O. Verscheure, P. Frossard, and M. Hamdi, "User-oriented QoS analysis in MPEG-2 video delivery," *Real-Time Imag.*, vol. 5, p. 305–314, 1999.

[45] K. Yamagishi and T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services," in *Proc. IEEE ICC*, 2008, pp. 110–114.

[46] H. Koumaras, C.-H. Lin, C.-K. Shieh, and A. Kourtis, "A framework for end-to-end video quality prediction of MPEG video," *J. Vis. Commun. Image Rep.*, 2009.

[47] F. You, W. Zhang, and J. Xiao, "Packet loss pattern and parametric video quality model for IPTV," in *Proc. IEEE/ACIS ICIS*, 2009, pp. 824–828.

[48] M.-N. Garcia and A. Raake, "Parametric packet-layer video quality model for IPTV," in *Proc. ISSPA*, 2010, pp. 349–352.

[49] T. Yamada, S. Yachida, Y. Senda, and M. Serizawa, "Accurate video-quality estimation without video decoding," in *Proc. IEEE ICASSP*, 2010, pp. 2426–2429.

[50] J. Gustafsson, G. Heikkilä, and M. Petterson, "Measuring multimedia quality in mobile networks with an objective parametric model," in *Proc. IEEE ICIP*, 2008, pp. 405–408.

[51] O. Sugimoto, S. Naito, S. Sakazawa, and A. Koike, "Objective perceptual video quality measurement method based on hybrid no reference framework," in *Proc. IEEE ICIP*, 2009, pp. 2237–2240.

[52] T. Brandao and M. P. Queluz, "No-reference quality assessment of H.264/AVC encoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1437–1447, 2010.

[53] M.-N. Garcia, R. Schleicher, and A. Raake, "Towards a content-based parametric video qualiy model for IPTV," in *Proc. VPQM*, 2010, pp. 20–25.

[54] K. Yamagishi, T. Kawano, and T. Hayashi, "Hybrid video-quality-estimation model for IPTV services," in *Proc. IEEE Globecom*, 2009, pp. 1–5.

[55] S. Kanumuri, S. Subramanian, P. Cosman, and A. Reibman, "Predicting H.264 packet loss visibility using a generalized linear model," in *Proc. IEEE ICIP*, 2006, pp. 2017–2020.

[56] N. Suresh and N. Jayant, "AVQ: A zero-reference metric for automatic measurement of the quality of visual communications," in *Proc. VPQM*, 2007, pp. 1–4.

[57] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for H.264/AVC coded video," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 932–946, 2009.

[58] ITU, "Opinion model for video-telephony applications," *ITU-T Rec. G.1070, Int. Telecomm. Union*, Geneva, Switzerland, 2007.

[59] M. Barkowsky, R. Bitto, J. Bialkowski, and A. Kaup, "Comparison of matching strategies for temporal frame registration in the perceptual evaluation of video," in *Proc. VPQM*, 2006, pp. 1–4.

[60] A. Clark, "Method and system for viewer quality estimation of packet video streams," U.S. Patent 2009/0 041 114 A1, 2009.

[61] K. Watanabe, K. Yamagishi, J. Okamoto, and A. Takahashi , "Proposal of new QoE assessment approach for quality management of IPTV services," in *Proc. IEEE ICIP*, 2008, pp. 2060–2063.

[62] N. Suresh and N. Jayant, "'Mean time between failures': A subjectively meaningful video quality metric," in *Proc. IEEE ICASSP*, vol. 2, 2006, pp. 1–4.

[63] Q. Li and Z. Wang, "Video quality assessment by incorporating a motion perception model," in *Proc. IEEE ICIP*, 2007, pp. 173–176.

[64] J. You, J. Korhonen, and A. Perkis, "Attention modeling for video quality assessment: Balancing global quality and local quality," in *Proc. IEEE ICME*, 2010, pp. 914–919.

[65] Video Quality Experts Group (VQEG). [Online]. Available: www.vqeg.org

[66] Joint Effort Group (JEG). [Online]. Available: http://wiki.vqeg-jeg.org/

[67] W. Chen, J. Fournier, M. Barkowski, and P. L. Callet, "New stereoscopic video shooting rule based on stereoscopic distortion parameters and comfortable viewing zone," in *Proc. SPIE*, vol. 7863, 2011, p. 78631O.

[68] G. Akar, M. Tekalp, C. Fehn, and R. Civanlar, "Transport methods in 3DTV—A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1622–1630, 2007.

[69] P. Campisi, P. L. Callet, and E. Marini, "Stereoscopic images quality assessment," in *Proc. EUSIPCO*, 2007, pp. 2110–2114.

[70] D. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, 2004.

[71] BBF, "Data model for a TR-069 enabled STB," *Broadband Forum, TR-135*, 2007.

[72] IETF, "RTP control protocol extended reports (RTCP XR)," *Internet Eng. Task Force, IETF RFC 3611*, 2003.

[73] BBF, "CPE WAN management protocol (CWMP)," *TR-069, Broadband Forum*, Nov. 2010.

[74] 3GPP, "Transparent end-to-end packet-switched streaming service (PSS); Protocols and codecs," *3GPP TS 26.234*, June 2011, v10.1.0.

[75] 3GPP, "IP multimedia subsystem (IMS); Multimedia telephony; Media handling and interaction," *3GPP TS 26.114*, June 2011, v11.0.0.

[76] OMA, "Enabler release definition for OMA device management," *OMA-ERELD-DM-V1_2-20070209-A, Open Mobile Alliance,* 2007, appr. Vers. 1.2.

[77] N. Staelens, S. Moens, W. Van den Broeck, I. Mariën, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of iptv and video on demand services in real-life environments," *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 458–466, 2010.

[78] M. Lambooij, I. Wijnand, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *J. Imag. Sci. Technol.*, vol. 53, no. 4, pp. 030 201–030 201-14, 2009.

[79] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 968–976, 2005.

[80] F. Rumsey, S. Zielinski, P. Jackson, M. Dewhirst, R. Conetta, S. George, S. Bech, and D. Meares "QESTRAL (Part 1): Quality evaluation of spatial transmission and reproduction using an artificial listener," in *Proc. 125th AES Conv.*, 2008.

[SP]

[Abdullah Bulbul, Tolga Capin, Guillaume Lavoué, and Marius Preda]

Assessing Visual Quality of 3-D Polygonal Models

[Evaluating and discussing existing metric performance]

Recent advances in evaluating and measuring the perceived visual quality of three-dimensional (3-D) polygonal models are presented in this article, which analyzes the general process of objective quality assessment metrics and subjective user evaluation methods and presents a taxonomy of existing solutions. Simple geometric error computed directly on the 3-D models does not necessarily reflect the perceived visual quality; therefore, integrating perceptual issues for 3-D quality assessment is of great significance. This article discusses existing metrics, including perceptually based ones, computed either on 3-D data or on two-dimensional (2-D) projections, and evaluates their performance for their correlation with existing subjective studies.

Digital Object Identifier 10.1109/MSP.2011.942466
Date of publication: 1 November 2011

INTRODUCTION

Technologies underlying 3-D computer graphics have matured to the point that they are widely used in several mass-market applications, including networked 3-D games, 3-D virtual and immersive worlds, and 3-D visualization applications [1]. Furthermore, emerging products, such as 3-D TVs and 3-D-enabled gaming devices, are opening new avenues of opportunity for an enhanced user experience when interacting with 3-D environments. Thus, 3-D models are emerging as a newly popular form of media [2], usually in the form of 3-D polygonal meshes.

1053-5888/11/$26.00©2011IEEE

Three-dimensional mesh models are generally composed of a large set of connected vertices and faces required to be rendered and/or streamed in real time. Using a high number of vertices/faces enables a more detailed representation of a model and possibly increases the visual quality while causing a performance loss because of the increased computations. Therefore, a tradeoff often emerges between the visual quality of the graphical models and processing time, which results in a need to judge the quality of 3-D graphical content. Several operations in 3-D models need quality evaluation. For example, transmission of 3-D models in network-based applications requires 3-D model compression and streaming, in which a tradeoff must be made between the visual quality and the transmission speed. Several applications require accurate level-of-detail (LOD) simplification of 3-D meshes for fast processing and rendering optimization. Watermarking of 3-D models requires evaluation of quality due to artifacts produced. Indexing and retrieval of 3-D models require metrics for judging the quality of 3-D models that are indexed. Most of these operations cause certain modifications to the 3-D shape (see Figure 1). For example, compression and watermarking schemes may introduce aliasing or even more complex artifacts; LOD simplification and denoising result in a kind of smoothing of the input mesh and can also produce unwanted sharp features. To bring 3-D graphics to the masses with a high fidelity, different aspects of the quality of the user experience must be understood.

SIMPLE GEOMETRIC ERROR COMPUTED DIRECTLY ON THE 3-D MODEL DOES NOT NECESSARILY REFLECT THE PERCEIVED VISUAL QUALITY.

Three-dimensional mesh models, as a form of visual media, potentially benefit from well-established 2-D image and video assessment methods, such as the visible difference predictor (VDP) [3]. Various metrics have thus been proposed that extend the 2-D objective quality assessment techniques to incorporate 3-D graphical mechanisms. Several aspects of 3-D graphics make them a special case, however. Three-dimensional models can be viewed from different viewpoints, thus, depending on the application, view-dependent or view-independent techniques may be needed. In addition, once the models are created, their appearance does not depend only on the geometry but also on the material properties, the texture, and the lighting [4]. Furthermore, certain operations on the input 3-D model, such as simplification, reduce the number of vertices; and this makes it necessary to handle changes in the input model.

VIEWPOINT-INDEPENDENT QUALITY ASSESSMENT

One category of quality assessment metrics directly works on the 3-D object space. The quality of a processed (simplified, smoothed, watermarked, etc.) model is generally measured in terms of how "similar" it is to a given original mesh. These similarity metrics measure the impact of the operations on the model. Another possible approach to evaluate the 3-D models is to consider 2-D rendered images of them according to certain viewpoints; however, viewpoint-independent error metrics would be necessary because they provide a unique quality value for a model even if it has been rendered from various viewpoints. Such metrics can be used for comparing compressed models or selecting a level of detail, for example.

GEOMETRIC-DISTANCE-BASED METRICS

The simplest estimation of how similar two meshes are is provided by the root mean square (RMS) difference

$$\mathrm{RMS}(A,B)=\sqrt{\sum_{i=1}^{n}\|a_i-b_i\|^2}, \tag{1}$$

where A and B are two meshes with the same connectivity, a_i and b_i are the corresponding vertices of A and B, and $\|\cdot\cdot\|$ is the Euclidean distance between two points. The problem is that this metric is limited to comparing meshes with the same number of vertices and connectivity.

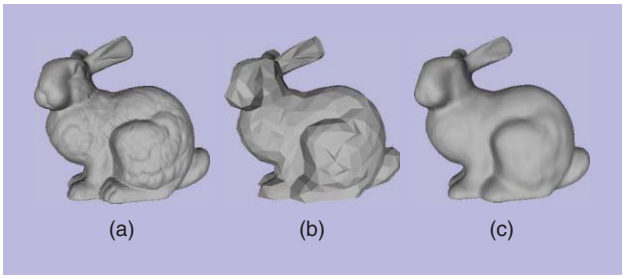
One of the most popular and earliest metrics for comparing a pair of models with different connectivities is the Hausdorff distance [5]. This metric calculates the similarity of two point sets by computing one-sided distances. The one-sided distance $D(A,B)$ of surface A to surface B is computed as follows:

$$\begin{aligned}\mathrm{dist}(a,B)&=\min_{b\in B}(\|a-b\|)\\ D(A,B)&=\max_{a\in A}(\mathrm{dist}(a,B)).\end{aligned} \tag{2}$$

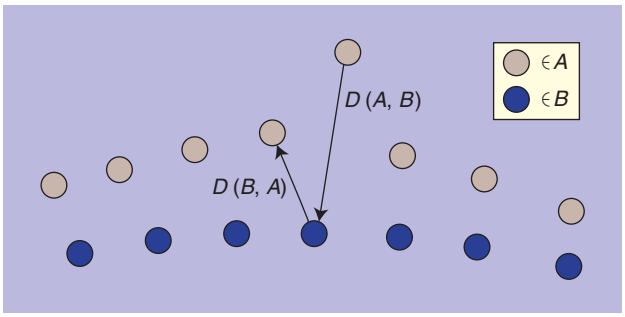
As this distance is nonsymmetric, the two-sided Hausdorff distance is computed by taking the maximum of $D(A,B)$ and $D(B,A)$ (Figure 2)

$$H(A,B)=\max(D(A,B),D(B,A)). \tag{3}$$

The Hausdorff distance has been used to find the geometric error between a pair of 3-D mesh models in the Metro tool by Cignoni et al. [5]. In this approach, the mean distance between a pair of meshes is found by dividing the surface integral of the distance between the two meshes by the area of one of the surfaces. The computation of this integral on discrete 3-D models



[FIG1] The bunny model: (a) original, (b) simplified, and (c) smoothed. (Used with permission from the Stanford University Computer Graphics Laboratory.)



[FIG2] The Hausdorff distance between two surfaces. The two-sided Hausdorff distance is $H(A, B) = \max(D(A,B), D(B, A))$ [6].

requires a sampling method for fast computation. Aspert et al. [7] also propose a sampling implementation of the Hausdorff distance in the MESH tool.

The Hausdorff distance computes the final distance between two surfaces as the maximum of all pointwise distances. Rather than taking the maximum, extensions have been proposed to provide a better indication of the error across the entire surface. Instead of taking the maximum of the pointwise distances, the average (known as the L_1 norm), the RMS (L_2 norm), and combinations have been proposed [5], [6].

These metrics are well known and widely used; however, even if they can correctly correlate with human judgement in some simple scenarios (see the section "Subjective Evaluation of 3-D Polygonal Models"), they usually fail to reflect the perceived quality because they compute a pure geometric distance between a pair of meshes, ignoring the working principles of the human visual system. Hence, several other metrics, using different perceptual principles, have been proposed to better estimate the perceived quality of 3-D meshes. These solutions can be categorized as roughness-based, structure-based, saliency-based, and strain-energy-based metrics. Since each of these categories focuses on different aspects of perception, it is unlikely for one of them to estimate the perceived visual quality for all scenarios. In this case, blending metrics of several categories may be a possible solution.

ROUGHNESS-BASED METRICS

Several solutions evaluate the quality of processed 3-D models based on their differences from the original model in their surface roughness (or smoothness). These solutions employ the observation that operations on 3-D mesh either introduce a kind of noise related to roughness (e.g., as with quantization or watermarking) or cause smoothing of the surface details (e.g., with LOD simplification for rendering). Roughness is an important perceptual property, as we cannot determine the effect of a small distortion if it is on a rough region of the model, and we can detect defects on smooth surfaces more easily. This perceptual attribute, called the masking effect, states that one visual pattern can hide the visibility of another.

Karni and Gotsman propose such a roughness-based error metric to evaluate their mesh compression approach [9]. This metric calculates the geometric Laplacian (GL) of a vertex v_i as follows:

$$\mathrm{GL}(v_i) = v_i - \frac{\sum_{j \in n(i)} l_{ij}^{-1} v_j}{\sum_{j \in n(i)} l_{ij}^{-1}}, \tag{4}$$

where $n(i)$ is the set of neighbors of vertex i, and l_{ij} is the geometric distance between vertices i and j. Then the norm of the Laplacian difference between models M^1 and M^2 is combined with the norm of the geometric distance between the models as follows (v is the vertex set of M)

$$\|M^1 - M^2\| = \frac{1}{2n}(\|v^1 - v^2\| + \|GL(v^1) - GL(v^2)\|). \tag{5}$$

One limitation of this metric is that the compared models must have the same connectivity as the RMS error approach.

Wu et al. [10], for driving their simplification algorithm, examine the dihedral angles of the adjacent faces, considering that a rough surface should have greater dihedral angles. Roughness variation has also been used for quality assessment of watermarked meshes; Gelasca et al. [11] and Corsini et al. [12] measure roughness strength by taking the difference between a mesh and its smoothed version. After computing roughness values for the original and watermarked models, the roughness-based difference is calculated as follows:

$$R(M, M^w) = \log\left(\frac{R(M) - R(M^w)}{R(M)} + k\right) - \log(k), \tag{6}$$

where $R(M)$ is the roughness of the original mesh, $R(M^w)$ is the roughness of the watermarked mesh, and k is a constant to stabilize the numerical results. These roughness-based perceptual metrics [11], [12] have shown to correlate very well with human judgement, particularly in the context of watermarking distortions.

Lavoué proposes a local roughness measure that is able to efficiently differentiate between the different kinds of regions in a mesh: rough parts, smooth regions, and "edge" features, which define border areas between regions [13] (see Figure 3). The proposed measure is based on a curvature analysis of local windows of the mesh and is independent of its connectivity. This measure does not estimate any distance but provides a local roughness estimation that can be used to hide artifacts and could be useful for the design of future quality metrics.

STRUCTURAL DISTORTION-BASED METRICS

Structural distortion-based metrics consider the assumption that the human visual system is good at extracting the

structural information of a scene in addition to local properties. Lavoué et al. [14] propose mesh structural distortion measure (MSDM), based on the work of Wang et al. [15], dedicated to 2-D images. Instead of extracting the structural information using luminance in 2-D images, this metric uses curvature analysis of the mesh geometry. In this work, a local MSDM (LMSDM) on two local windows x and y of the two meshes is calculated as

EMERGING AS A NEWLY POPULAR FORM OF MEDIA, 3-D MODELS ARE USUALLY IN THE FORM OF 3-D POLYGONAL MESHES.

$$\mathrm{LMSDM}(x,y) = \big(\alpha \times L(x,y)^a + \beta \times C(x,y)^a + \gamma \times S(x,y)^a\big)^{\frac{1}{a}}, \quad (7)$$

with α, β, and γ selected as 0.4, 0.4, and 0.2, respectively, by the authors and with curvature comparison L, contrast comparison C, and structure comparison S computed as

$$L(x,y) = \frac{\|\mu_x - \mu_y\|}{\max(\mu_x,\mu_y)},$$
$$C(x,y) = \frac{\|\sigma_x - \sigma_y\|}{\max(\sigma_x,\sigma_y)}, \quad \text{and}$$
$$S(x,y) = \frac{\|\sigma_x\sigma_y - \sigma_{xy}\|}{\sigma_x,\sigma_y}, \quad (8)$$

where μ_x, σ_x, and σ_{xy} are respectively the mean, standard deviation, and covariance of the curvature on local windows x and y. Then the MSDM is calculated as follows:

$$\mathrm{MSDM}(X,Y) = \left(\frac{1}{n_w}\sum_{i=1}^{n_w}\mathrm{LMSDM}(x_i,y_i)^a\right)^{\frac{1}{a}} \in [0,1), \quad (9)$$

where X and Y are the compared meshes, x_i and y_i are the corresponding local windows of the meshes, and n_w is the number of local windows. The value a is selected as three by the authors, for (7) and (9) [14]. This metric has proven to correlate very well with human judgement even in difficult scenarios. The authors propose an improved version of this method in [16].

SALIENCY-BASED METRICS

The metrics described above provide a guarantee of the maximum geometric distance rather than estimating the perceived distance between the models.

In this group of metrics, the idea is to give more importance to parts of the meshes that gather more human attention. This type of metric is generally used for mesh simplification such that salient parts of a mesh are preserved in the simplification, as suggested by Howlett et al. [17] and Lee et al. [18]. The salient parts of meshes are determined by utilizing an eye-tracker in Howlett et al.'s work, whereas Lee et al.'s method is more convenient as it computes saliency of a mesh automatically, based on its surface curvature.

Similar to the roughness-based and structural distortion-based metrics, saliency uses the perceptual limitation of the human visual system, and its further use for mesh quality assessment is a research area of great interest.

STRAIN-ENERGY-BASED METRICS

Bian et al. [19] propose a solution based on the strain energy on the mesh as a result of elastic deformation. Mesh models are assumed to be elastic objects; as shells composed of triangular faces of negligible thickness. The assumption is that triangle faces do not bend, and each triangle is deformed along its plane by ignoring any rigid body motion.

The perceptual distance between the two versions of the input model is defined as the weighted average strain energy (ASE) over all triangles of the mesh, normalized by the total area of the triangular faces

$$\mathrm{SFEM}(A,B) = \frac{1}{S}\sum w_i W_i, \quad (10)$$

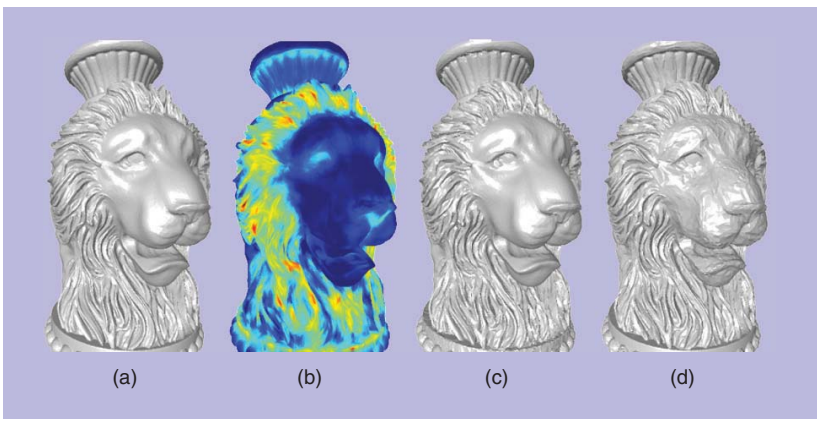
where w_i are weights for which several strategies are tested in [19] and W_i is the strain energy computed for triangle i.

This model correlates well with human opinion from the subjective experiment conducted by the authors.

ATTRIBUTE-BASED METRICS

Many 3-D mesh models contain per-vertex attributes in addition to the vertex position, such as color, normal, and texture coordinates. Also, in sharp creases of the models, there may be multiple normals per-vertex, or there may be several color values on the boundaries, causing discontinuities in the attributes.

As described by Luebke et al. [6], correspondence between vertices on two surfaces is important but is a



[FIG3] Roughness map of a 3-D model. (a) Original model, roughness map; (b) rough regions shown with warmer colors; (c) noise on rough regions; and (d) noise on smooth regions [8].

difficult issue for meshes with different connectivities; it is difficult to compare attribute values from the original surface and a simplified version in a continuous function. Luebke describes an alternative to Hausdorff, called the bijection method. This requires correspondence between vertices in a 2-D parametric domain, such as a texture map. This distance is called a parametric distance. Roy et al. [20] propose a metric called attribute deviation metric, that can be used to compare two meshes according to their geometric and appearance attributes (or any other per-vertex attributes). The local deviation of attributes between each point of a mesh and the surface of the reference mesh is calculated using parametric distances.

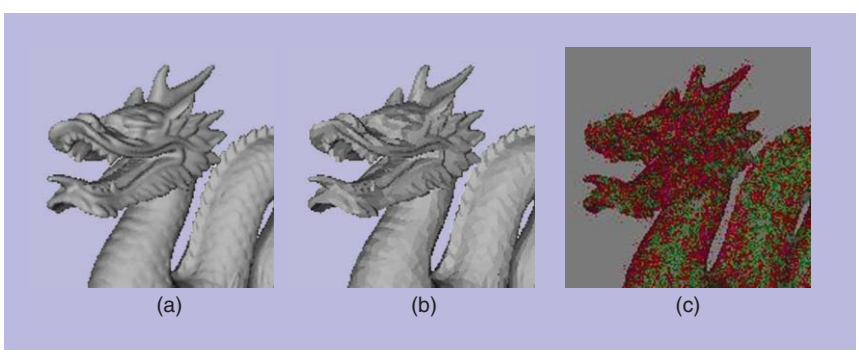
Pan et al. propose a different approach for quality assessment, calculating the quality of a 3-D model according to its wireframe and texture resolutions (11) [21]

$$Q(g,t) = \frac{1}{\frac{1}{m+(M-m)t} + \left(\frac{1}{m} - \frac{1}{m+(M-m)t}\right)(1-g)^c}. \tag{11}$$

Here, m and M are the minimum and maximum bounds of quality, g and t are graphical and texture components scaled into a $[0-1]$ interval, and c is a constant. All coefficients are determined by curve fitting on subjective evaluation data. This metric provides a very good estimation of human judgement as demonstrated in the authors' subjective experiment.

VIEWPOINT-DEPENDENT QUALITY ASSESSMENT

Viewpoint-dependent quality assessment metrics estimate the perceptual quality of a 3-D model as it is shown on the screen; therefore, these metrics are image based. Viewpoint-dependent metrics can be classified as nonperceptual metrics and perceptually based metrics. The visual system does not matter for nonperceptual approaches; they compute the difference between two images pixel by pixel. Perceptually based metrics rely on the mechanisms of the human visual system and attempt to predict the probability that the human observer will be able to notice differences between images.



[FIG4] (a) Original image, (b) simplified image, and (c) VDP output. (Used with permission from the Stanford University Computer Graphics Laboratory.)

NONPERCEPTUAL METRICS

Lindstrom and Turk calculate the RMS image error for mesh simplification [22]. In their work, the meshes are rendered from multiple viewpoints and the quality of the resulting luminance images are measured in terms of their differences from the original image as follows:

$$d_{\mathrm{RMS}}(Y^0,Y^1) = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(y_{ij}^0 - y_{ij}^1\right)^2}, \tag{12}$$

where Y^0 and Y^1 are m by n luminance images. The RMS metric is not a good metric for image quality assessment and is seldom used because it is highly affected by a shift or scale, and it does not have a perceptual aspect.

Another quality metric for comparing image quality against a reference image consists in calculating the peak signal-to-noise ratio (PSNR). Using the RMS error shown in (12), the PSNR for an image with a highest possible intensity value $I_{\max}$ can be calculated by

$$\mathrm{PSNR} = 20\log_{10}\left(\frac{I_{\max}}{d_{\mathrm{RMS}}}\right). \tag{13}$$

Although PSNR is also widely used for natural images, it is shown to be a poor indicator of image quality [23]. However, according to a report of the Video Quality Experts Group (VQEG), many more complicated image quality metrics are not significantly better than PSNR [24]. The reasons for this are discussed in a study of Wang et al. [25].

PERCEPTUALLY BASED METRICS

Many 2-D metrics incorporate the mechanisms of the human visual system. These metrics generally use the following perceptual concepts: contrast sensitivity function (CSF), which indicates the relation between the visible spatial frequency and different contrast values; and masking, which describes the reduction in the visual sensitivity of a signal upon the existence of another signal.

A popular metric in this category is Daly's VDP [3]. This metric takes two images as inputs, one of which is evaluated relative to the other; and the output is an image of the perceptual differences between the two images (see Figure 4). The value of each pixel on the output image indicates the detection probability of the difference. The VDP is shown to be a good indicator of perceptually important areas in 3-D graphics scenes by the psychophysical experiment of Longhurst and Chalmers [26].

Another well-known metric is the Sarnoff visual discrimination model (VDM) [27] by Lubin. This metric also predicts the detection probability of the differences between a reference image and the evaluated image, as in VDP. The Sarnoff VDM model works on

spatial domain whereas VDP works in the frequency domain; VDM works faster but requires more memory. Li et al. [28] compare the two metrics and find that each model has advantageous properties.

Bolin and Meyer modify the Sarnoff VDM model and propose a simpler and faster metric, which incorporates color properties into their 3-D global illumination calculations [29]. This metric is preferred for its efficiency. In their subjective experiment with differently simplified 3-D models, Watson et al. [30] show that this metric is an effective predictor of fidelity.

Ramasubramanian et al. [31] propose a perceptually based metric that defines a threshold map in which the minimum detectable difference values are stored for each pixel. This metric handles luminance and spatial processing separately, which provides efficiency since it enables precomputing of the spatial features.

Ramanarayanan et al. [32] introduce a novel concept, the visual equivalence predictor (VEP), which claims that two images are visually equivalent if they give the same impression even though they have visually different parts. This concept makes more sense for computer-generated imagery in which slightly different illumination techniques lead to different images when analyzed pixelwise although the two images have similar fidelity and information. This model takes 3-D geometry, material, and illumination properties into account for the equivalency computations. The VEP concept aims to overcome the limitations of the VDP model, which only considers the earliest levels of visual coding, and is therefore too conservative with respect to the kinds of approximations that can be applied in the rendering process.

Visual masking, which describes the reduction in visual sensitivity of a signal upon the existence of another signal, has been used for view-dependent quality assessment of 3-D models. Ferwerda et al. [33] investigate the masking effect for computer graphics and extend the VDP model to include color. In their study, a computational model of the masking effect of the used textures on the artifacts of the 3-D meshes is developed. This masking effect is predicted on the varying contrast, spatial frequency, and orientation features of the texture pattern and on the polygonal tessellation of the model surface.

Three image-quality metrics based on perceptual color differences are proposed by Albin et al. [34]. These similar metrics find the difference between two images in the LLAB (a modified version of CIELAB) color space. The authors state that these metrics are not complete but only initial attempts at a perceptual quality metric. While the first metric is based on a pixel-by-pixel difference of the images, the second metric gives a single distance value using a Monte-Carlo approach, and the last one is a subdivision-based metric, which gives a rougher difference image compared to the first metric in a shorter time.

THE PARAMETERS USED IN AN EXPERIMENT ARE OF GREAT IMPORTANCE BECAUSE THEY CAN BIAS THE RESULTS SIGNIFICANTLY, ESPECIALLY FOR COMPUTER-GENERATED STIMULI.

SUBJECTIVE EVALUATION OF 3-D POLYGONAL MODELS

While automatic metrics are commonly used to predict perceptual quality, relatively few researchers have attempted to measure and predict the visual fidelity of 3-D models through subjective experiments. These experiments could be directly used to predict the perceptual quality of 3-D models as well as to validate the outcomes of automatic metrics described in the previous sections. Generally, the term "quality" is used to judge how two images (one of them original, the other modified) are "similar" to each other.

EXPERIMENTAL MEASURES

Watson et al. study experimental fidelity measures for 3-D graphical models [30] and define three of them:

1) naming time, which measures the time from the appearance of an object until the observer names it
2) rating, where observers assign a number within a range and meaning determined by the experimenter
3) forced choice preferences, where observers are shown two or more stimuli, and they choose the stimulus with more of the experimenter-defined quality.

The results of this work show that automatic measures of fidelity (e.g., Bolin's [29], Metro [5], mean squared error (MSE) approaches) are successful at predicting experimental ratings, less successful at predicting preferences, and largely unsuccessful at predicting naming times. On the other hand, when the task is based on comparing different models, ranking is stated to be better than rating the models because the given ratings do not necessarily reflect the perceptual distance between the compared models [35], [36]. The experimental measures used in several user studies can be found in Table 1.

EXPERIMENTAL DESIGN

The parameters used in an experiment are of great importance because they can bias the results significantly, especially for computer-generated stimuli, where almost everything can be controlled. Effective parameters controlled in several quality assessment studies are shown in Table 2 and listed as follows:

- *Lighting:* The position and type of light source is a crucial element, with a major effect on the viewing conditions. Rogowitz et al. [35] show that models lit from the front result in different subjective scores compared to the same models lit from above. The human visual system has a prior that light is stationary and comes from a left-above orientation [37].
- *Background:* The background may affect the perceived quality by changing the visibility of the boundaries of the model. While a uniform black background is used in several user studies [30], [38], Corsini et al. [12] choose a nonuniform background that fades from blue to white so as not to overestimate the contours.

[TABLE 1] EXPERIMENT METHODOLOGIES OF RECENT SUBJECTIVE EXPERIMENTS ON QUALITY ASSESSMENT.

	MASKING	TASK	MEASURES	LEVELS	PURPOSE OF TEST
WATSON01		MESH SIMP.	RATING, PREFERENCE, NAMING TIME	3 (20%, 50%, ORIG)	
ROGOWITZ01		MESH SIMP.	RATING	3 (25%, 40%, ORIG)	TO EVALUATE STILL IMAGES FOR GEO. MODELS
CORSINI07		WATERMARKING	RATING	E1: 4 MOD × 3 WM LEVELS × 3 RES. + 4 ORIG E2: 4 MOD × 11 WM TYPES + 4 ORIG	TO FIT A METRIC
SILVA08		QUAL. ASSESMENT ON MESH SIMP.	RANKING 4 MODELS	10, 20, 27, 35, 43, 50%	TO CALCULATE QUALITY OF SIMPLIFIED MESHES
PAN05	TEXTURE-MESH GEOMETRY	FITTING OF SUBJECT EVAL. RESULTS TO PERCEPTUAL METRIC	RATING – 5 LEVELS	5 OBJECTS, 6 LEVELS OF MESH RESOLUTION × 3 LEVELS OF TEXTURE RESOLUTION	TO FIT A METRIC
LAVOUÉ10	NOISE-ROUGHNESS	COMPRESSION, WATERMARKING, SMOOTHING	RATING	E1: ORIG + 3X NXOISE (0.02, 0.01, 0.005 OF BBOX) E2: 4 ORIG + 4×3 NOISE ON SMOOTH + 4×3 NOISE ON ROUGH E3: 4 ORIG + 4×9 SMOOTH + 4×12 NOISE VERSIONS	TO COMPARE DIFFERENT OBJECTIVE METRICS
RUSHMEIER00	TEXTURE-MESH GEOMETRY	SIMPLIFICATION	RATING: 0...100	2 OBJECTS × 3 GEOMETRY LEVELS (FULL; 47X REDUCTION IN SIZE; 94X REDUCTION IN SIZE) × 4 TEXTURE LEVELS (NONE, 512×512, 256×256, 64×64)	TO EXAMINE POSSIBILITY OF SUBSTITUTING TEXTURE FOR GEOMETRY

■ *Materials and shading:* Today, almost all 3-D models used in applications have material properties (e.g., texture, normals) and associated complex programmable shaders. On the other hand, most of the subjective evaluations for verifying perceptual metrics do not take material properties into account; they use only diffuse and smooth-shaded models, mostly to prevent highlight effects [12]. Textures have only been used in the context of substituting geometry with texture [21], [38]. On the other hand, as described above, material properties such as textures introduce the masking effect and hide visual artifacts. Researchers often use models without textures or complex material properties to better control the number of variables influencing the outputs.

■ *Animation and interaction:* To evaluate a 3-D model in a fair way, observers should be able to see the models from different viewpoints. This can be achieved by animating the object or viewpoint as in [21], [35], as well as giving free viewpoint control to the user as in [8], [12], and [36]. Furthermore, animations affect the perception of the models such that, in the study of Rogowitz and Rushmeier [35], artifacts caused by simplification are less visible when the objects are rotating rather than standing still. The sensitivity of the human visual system is dependent on retinal velocity; the eye's tracking ability is limited to 80 °/s [39], which should be taken into account when an experiment includes animation.

■ *Type of objects:* There are several concerns to keep in mind when selecting objects for a subjective experiment. Watson et al. [30] state that evaluation results are different for animal models and man-made artifacts. Further, using abstract objects helps avoid semantic interpretation [38]. Also, the complexity and roughness of the models are important. In a very complex object, simplifications may not be visible and the roughness of a mesh may mask artifacts.

■ *Masking:* The object's geometry, roughness, texture, and applied noise or watermarking can mask each other. Lavoue et al. [13] examine the masking effect of noise and roughness; Pan et al. [21] and Rushmeier et al. [38] examine the masking effect of textures on geometry. The masking effect should be considered while designing an experiment.

■ *Extent:* The extent, i.e., the display area of the rendered model in pixels, should be large enough to reflect the details of the model. Showing too many items simultaneously may decrease the visibility of the models. The display extents used in several user studies can be found in Table 2.

■ *Levels:* When an operation (simplification, watermarking, etc.) on meshes is to be tested, the number of the comparison cases and the strengths of the applied operations for each case should be adjusted carefully. Too few levels (compared cases) may not sufficiently reflect the tested operation, whereas a large number of levels may not be feasible, as they would require too many subjects. For simplification case, there are studies using three [30], [35] to seven [36] levels (including the originals) of simplification.

- *Stimuli order:* In comparison-based experiments, stimuli can be shown to the user simultaneously (e.g., side by side) or in succession (e.g., first the reference, then the tested models). When they are shown in succession, enabling users to turn back to the reference model as in the experiment of Rogowitz and Rushmeier [35], allows for a more detailed comparison. Also, the order and the position of the stimuli should be selected in a way that minimizes the effect of external variables such as observer movements and room's ambient light.
- *Duration*: The duration of which the tested models are shown to the subjects may also affect the results of evaluation.

STANDARDS FOR SUBJECTIVE EVALUATION

Although no specific recommendation for subjective evaluation of 3-D models exists currently, a number of standards, which define the conditions for subjective experiments for other multimedia content (e.g., image and video), could be adapted and used. A well-known standard is the ITU-R BT.500 Recommendation [40], which defines the methodology for the subjective evaluation of image quality. Different experiment methods, such as double-stimulus continuous quality-scale (DSCQS) and simultaneous double stimulus for continuous evaluation (SDSCE) are recommended and grading scales and how to present test materials are outlined. Several of these methods, which may be useful for quality assessment of 3-D meshes, are briefly explained below.

- The DSCQS method is recommended for measuring the relative quality of a system against a reference. It has a continuous grade scale that is partitioned into five divisions of equal length, labeled bad, poor, fair, good, and excellent. Subjects can mark the scale in a continuous manner and then the grades are mapped to a zero to 100 interval. The reference and test material are shown twice in succession.
- The SDSCE method is recommended for measuring the fidelity between two impaired video sequences. The stimuli are shown side by side and the grading is continuous.
- The ITU-R BT.500 standard also includes recommendations related to the evaluation of the experiments, such as how to eliminate the outlier data.

A related standard, the ITU-T P.910 recommendation [41], describes subjective assessment methods for evaluating the one-way overall quality for multimedia applications. This recommendation addresses test methods and experiment design, including comparison methods; and evaluation procedures, including viewing conditions and the characteristics of the source sequences, such as duration, kind of content, number of sequences, etc. Subjective evaluation of 3-D graphical models as a form of media can benefit from these recommendations.

PERFORMANCE EVALUATION

This section is an attempt to provide a performance comparison between the viewpoint-independent and viewpoint-dependent metrics. As an indicator of performance, we restrict to the

[TABLE 2] EXPERIMENT DESIGN OF RECENT SUBJECTIVE EXPERIMENTS ON QUALITY ASSESSMENT.

	LIGHTING	ANIMATION/ INTERACTION	MATERIALS	BACKGROUND	OBJECT	EXTENT OF STIMULUS	SIMULTANEOUS/ SUCCESSIVE
WATSON01 [30]	OBLIQUE	NO/NO	NO	BLACK	MAN-MADE VS. ANIMAL	591PX IN WIDTH	SIMULTANEOUS
ROGOWITZ01 [35]	ABOVE, COLOCATED WITH VIEW	ROTATING OBJECT/NO	NO		SIMPLE VS. COMPLEX (VERTEX COUNT)		SUCCESSIVE, CAN GO BACK
CORSINI07 [12]	WHITE POINT LIGHT, TOP CORNER OF OBJ. BBOX	NO/FREE INTERACTION	DIFFUSE ONLY	BLUE TO WHITE FADING		600×600PX	
SILVA08 [36]		NO/FREE INTERACTION		BLACK	HAVE SMALL NUM. OF VERTICES AND DIFFERENT NATURE		SIMULTANEOUS
PAN05 [21]	FRONT	ROTATION/ADJUSTABLE ROT. SPEED	TEXTURED OBJECTS			750PX IN HEIGHT	SIMULTANEOUS
LAVOUÉ10 E1, E2	FRONT	NO/FREE INTERACTION	DIFFUSE ONLY	BLACK	OBJECTS WITH SMOOTH AND ROUGH REGIONS		SIMULTANEOUS
LAVOUÉ10 E3 [8]	WHITE POINT LIGHT, TOP CORNER OF OBJ. BB.	NO/FREE INTERACTION	DIFFUSE ONLY	NONUNIFORM	OBJECTS FROM DIFFERENT NATURES		SUCCESSIVE
RUSHMEIER00 [38]	ABOVE, COLOCATED WITH VIEW	NO/NO	TEXTURED OBJECTS	BLACK	ABSTRACT OBJECTS	370PX IN HEIGHT	

correlation with human judgment (Mean opinion scores coming from the rating of distorted models) captured via the subjective experiments presented in the section "Subjective Evaluation of 3-D Polygonal Models."

THE USAGE OF TEXTURES, WHICH IS PRESENT IN ALMOST ALL CURRENT 3-D MODELS, HAS A GREAT IMPACT ON PERCEIVED QUALITY SINCE THEY CAN MASK ARTIFACTS ON THE MODELS.

The task of comparison is difficult because most of the metrics have been evaluated by a subjective database associated with their own protocols, which sometimes leads to contradictory results. For instance, Rogowitz et al. [35] state that image quality metrics are not adequate for measuring the quality of 3-D graphical models because lighting and animation affect the results; on the other hand, based on their experiment with models of varying LODs, Cleju and Saupe [42] claim that image-based metrics are better than metrics working on 3-D geometry.

Table 3 presents the performance evaluation for most of the existing metrics according to the difficulty of the database; indeed if the stimuli come from the same source of distortion (such as a uniform noise addition) even a simple metric will correlate. On the other hand, if the stimuli come from different types of distortions applied on very different 3-D models, then it becomes much more difficult to correlate with human judgement. These data from Table 3 come from many different sources; for example, in the simplification column, we have synthesized results from Cleju and Saupe [42], Watson et al. [30], and Lavoué [16] using a database from Silva et al. [36]. Most of these results also come from the recent study of Lavoué and Corsini [8], who made a quantitative comparison of several 3-D perceptual metrics according to several subjective databases. The rating ranges from "–" for a poor correlation to "++" for an excellent one. To indicate that no data is available, "?" is used, and "N/A" stands for "not applicable."

Several conclusions can be drawn from this table: most metrics are still not applicable for evaluating simplification because they are not able to compare meshes that do not share the same connectivity [9], [14], [19] or the same sampling density [11], [12].

For simple scenarios (a single mesh processing method or a single model), even simple metrics (e.g., Hausdorff) are able to correlate well with human judgment. However, for more complex scenarios, the roughness-based and structural approaches [11], [12], [14], [16] largely perform better than these simple approaches.

Two-dimensional metrics seem to perform very well for evaluating simplification, but these metrics have never been tested on other scenarios. Actually, a quantitative performance comparison of these image-based metrics against recent 3-D perceptually based ones [11], [12], [14], [16] is still missing. This would be a very interesting task involving any parameters: Which 2-D metrics would perform best? How would one choose the 2-D views of the 3-D models to feed these metrics and how would one combine the scores coming from different views into a single one for the whole 3-D model?

[TABLE 3] PERFORMANCES OF EXISTING METRICS.

METRIC/DATABASE	SIMPLIFICATION [16], [30], [36], [42]	UNIF. NOISE ADDITION [8], [19]	WATERMARKING [11], [12], [14], [19]	MASKING SCENARIO [14]	NONUNIF. NOISE + SMOOTHING [14]
HAUSDORFF BASED	+	+	+	–	–
RMS	N/A	+	?	–	–
GL [9]	N/A	+	?	–	–
ROUGHNESS BASED [11], [12]	N/A	++	++	–	+
MSDM [14]	N/A	++	++	+	++
MSDM2 [16]	++	++	?	++	++
STRAIN ENERGY [19]	N/A	+	+	–	–
2-D METRICS [15], [29]	++	?	?	?	?

DISCUSSION AND CONCLUSION

Three-dimensional graphical models are emerging as a newly popular form of media, and the importance of quality assessment of these models is only expected to grow. We have surveyed different metrics and approaches for evaluating the visual quality of 3-D polygonal models; however, there are several issues to consider in the future.

It is possible to assess the visual quality of 3-D models through subjective user tests or automatic measures of quality through view-dependent or view-independent metrics. Regarding the perceived quality directly, subjective evaluation has an advantage over metrics; however, applying user tests is not practical or costly for all applications. In certain applications, such as LOD rendering, quality also has significance for run time; therefore, there is always a need for better automatic metrics for 3-D polygonal models.

Objective metrics are categorized as viewpoint-independent metrics and viewpoint-dependent metrics. Most viewpoint-dependent metrics work on the image space instead of working on the 3-D geometry of the models. Although there are important research findings on this issue, we cannot clearly say that one group of metrics is superior to the other.

The nature of the models also affects the perceived quality. An important finding of Watson et al. [30] is that automatic measures predict the experimental ratings worse for animals (objects with smooth shapes) than man-made artifacts. Also, using familiar objects or abstract objects may change the results due to the semantic interpretations of the objects [38]. Furthermore, the usage of

textures, which is present in almost all current 3-D models, has a great impact on perceived quality since they can mask artifacts on the models [21], [38]. Animation also affects the perceived quality of 3-D graphical models [43], [44]. Taking the above factors into consideration, there is a need for more-comprehensive quality metrics that consider these different channels of the 3-D model.

ONE OF THE MOST POPULAR AND EARLIEST METRICS FOR COMPARING A PAIR OF MODELS WITH DIFFERENT CONNECTIVITIES IS THE HAUSDORFF DISTANCE.

Humans' visual system characteristics are considered in many of the metrics to better reflect the perceived quality. The adaptation ability of the visual system, the masking effect, and the contrast sensitivity function are the mostly used concepts. In addition to these visibility-related models, attention-oriented metrics, which deal with predicting the highest-attended locations of 3-D models, have the potential for further development. Although attention and saliency concepts are studied for 3-D models and used in several applications such as mesh simplification [17], [18], [45], there is a need for further work on developing attention-based quality metrics.

This article has also surveyed various subjective evaluation approaches for 3-D model quality assessment. The experiment design has a significant effect on the perceived quality because design decisions, such as the location of light sources, free interaction with the model, or the extent of the model on screen, bias the results. Even if the existing subjective studies seem to have produced relevant results with no error or bias in the protocols, there is a critical need for a real standard for testing 3-D graphical models.

Finally, we list some online platforms, tools, and repositories related to 3-D model quality assessment. These are generally specialized for a specific application area. For mesh compression, the MPEG 3-D Graphics (3DG) group has initiated an activity on scalable complexity mesh compression to merge the theoretical models and the content used in real applications. The group uses an online platform [46] to be able to deal with 3-D graphics objects with various characteristics. For mesh simplification, there are several available tools including QSlim [47] and MeshLab [48]. Another important application area is watermarking as stated by recent studies [11], [12], [14]. Wang et al. [49] have recently presented a benchmarking system for evaluating the 3-D mesh watermarking methods. Several repositories and tools, including PolyMeCo [50] and AIM@SHAPE [51], are constructed to ease the validation process of quality assessment metrics by providing tools and common comparison sets. There are also various free, general-purpose tools that provide 3-D model comparison (e.g., using geometric distance) such as Metro [5], MESH [7], and MeshDev [20]. These platforms and repositories can serve as a tool for future research in this field.

ACKNOWLEDGMENT

The authors would like to thank Rana Nelson for proofreading. This work is supported by the Scientific and Technical Research Council of Turkey (TUBITAK, project number 110E029). The bunny (Figure 1) and the dragon (Figure 4) models were obtained from Stanford University Computer Graphics Laboratory with permission.

AUTHORS

Abdullah Bulbul (bulbul@cs.bilkent.edu.tr) received his B.S. degree from the Department of Computer Engineering, Bilkent University (Ankara, Turkey) in 2007; currently he is a Ph.D. candidate in the same department. Between 2008 and 2010, he worked as a researcher for the All 3-D Imaging Phone (3-DPhone) project, which is a project funded by the Seventh Framework program of the European Union. His current research interests include the use of perceptual principles in computer graphics and mobile 3-D graphics.

Tolga Capin (tcapin@cs.bilkent.edu.tr) is an assistant professor in the Department of Computer Engineering at Bilkent University. He received his Ph.D. degree from Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland in 1998. He has authored more than 25 journal papers and book chapters, 40 conference papers, and a book. He has three patents and ten pending patent applications. His research interests include networked virtual environments, mobile graphics, computer animation, and human-computer interaction.

Guillaume Lavoué (glavoue@liris.cnrs.fr) received the engineering degree in signal processing and computer science from CPE Lyon (2002), the M.Sc. degree in image processing from the University Jean Monnet, St-Etienne (2002), and the Ph.D. degree in computer science from the University Claude Bernard, Lyon, France (2005). Following a postdoctoral fellowship at the Signal Processing Institute (EPFL) in Switzerland, he has been an associate professor at the French engineering university INSA of Lyon, in the LIRIS Laboratory (UMR 5205 CNRS), since 2006. His research interests include 3-D model analysis and processing including compression, watermarking, perception, and 2-D/3-D recognition.

Marius Preda (marius.preda@int-evry.fr) received an engineering degree in electronics from University Politehnica (Bucharest, Romania) in 1998, and a Ph.D. degree in mathematics and informatics from University Paris V–René Descartes (Paris, France) in 2001. He is currently an associate professor at Institut Telecom. His research interests include generic virtual character definition and animation, rendering, low bitrate compression and transmission of animation, and multimedia composition and standardization. He is the chair of the 3-D graphics subgroup of ISO's Moving Picture Experts Group and is involved

in several research projects at the institutional, national and European levels.

REFERENCES

[1] J. Cooperstock, "Multimodal telepresence systems," *IEEE Signal Processing Mag.*, vol. 28, no. 1, pp. 77–86, Jan. 2011.

[2] L. Daly and D. Brutzman, "X3D: Extensible 3D graphics standard [standards in a nutshell]," *IEEE Signal Processing Mag.,* vol. 24, no. 6, pp. 130–135, Nov. 2007.

[3] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA,: MIT Press, 1993, pp. 179–206.

[4] K. Myszkowski, T. Tawara, H. Akamine, and H.-P. Seidel, "Perception-guided global illumination solution for animation rendering," in *Proc. 28th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. New York: ACM, 2001, pp. 221–230.

[5] P. Cignoni, C. Rocchini, and R. Scopigno, "Metro: Measuring error on simplified surfaces," *Comput. Graph. Forum,* vol. 17, no. 2, pp. 167–174, 1998.

[6] D. Luebke, B. Watson, J. D. Cohen, M. Reddy, and A. Varshney, *Level of Detail for 3D Graphics*. New York: Elsevier Science, 2002.

[7] N. Aspert, D. Santa-Cruz, and T. Ebrahimi, "Mesh: Measuring errors between surfaces using the Hausdorff distance," in *Proc. 2002 IEEE Int. Conf. Multimedia and Expo (ICME '02)*, 2002, vol. 1, pp. 705–708.

[8] G. Lavoué and M. Corsini, "A comparison of perceptually based metrics for objective evaluation of geometry processing," *IEEE Trans. Multimedia,* vol. 12, no. 7, pp. 636–649, Nov. 2010.

[9] Z. Karni and C. Gotsman, "Spectral compression of mesh geometry," in *Proc. 27th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'00)*. New York: ACM/Addison-Wesley, 2000, pp. 279–286.

[10] J.-H. Wu, S.-M. Hu, J.-G. Sun, and C.-L. Tai, "An effective feature-preserving mesh simplification scheme based on face constriction," in *Proc. 9th Pacific Conf. Computer Graphics and Applications (PG'01)*. Washington, DC: IEEE Comput. Soc., 2001, pp. 12–21.

[11] E. D. Gelasca, T. Ebrahimi, M. Corsini, and M. Barni, "Objective evaluation of the perceptual quality of 3D watermarking," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2005, pp. 241–244.

[12] M. Corsini, E. D. Gelasca, T. Ebrahimi, and M. Barni, "Watermarked 3D mesh quality assessment," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 247–256, 2007.

[13] G. Lavoué, "A local roughness measure for 3d meshes and its application to visual masking," *ACM Trans. Appl. Percept.*, vol. 5, pp. 21:1–21:23, Feb. 2009.

[14] G. Lavoué, E. D. Gelasca, F. Dupont, A. Baskurt, and T. Ebrahimi, "Perceptually driven 3D distance metrics with application to watermarking," in *Proc. SPIE Applications of Digital Image Processing XXIX,* 2006, vol. 6312, pp. 63120L.1–63120L.12.

[15] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[16] G. Lavoué, "A multiscale metric for 3D mesh visual quality assessment," in *Computer Graphics Forum,* vol. 30, no. 5, pp. 1427–1437, 2011.

[17] S. Howlett, J. Hamill, and C. O'Sullivan, "An experimental approach to predicting saliency for simplified polygonal models," in *Proc. 1st Symp. Applied Perception in Graphics and Visualization (APGV'04)*. New York: ACM, 2004, pp. 57–64.

[18] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh saliency," in *Proc. ACM SIGGRAPH 2005 Papers (SIGGRAPH'05)*. New York: ACM, 2005, pp. 659–666.

[19] Z. Bian, S.-M. Hu, and R. R. Martin, "Evaluation for small visual difference between conforming meshes on strain field," *J. Comput. Sci. Technol.,* vol. 24, pp. 65–75, Jan. 2009.

[20] M. Roy, S. Foufou, and F. Truchetet, "Mesh comparison using attribute deviation metric," *J. Image Graph.*, vol. 4, no. 1, pp. 1–14, 2004.

[21] Y. Pan, I. Cheng, and A. Basu, "Quality metric for approximating subjective evaluation of 3-D objects," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 269–279, Apr. 2005.

[22] P. Lindstrom and G. Turk, "Image-driven simplification," *ACM Trans. Graph.*, vol. 19, pp. 204–241, July 2000.

[23] Z. Wang, H. Sheikh, and A. Bovik, "No-reference perceptual quality assessment of jpeg compressed images," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, 2002, vol. 1, pp. 477–480.

[24] A. M. Rohaly, J. Libert, P. Corriveau, and A. Webster, "Final report from the video quality experts group on the validation of objective models of video quality assessment," *Video Quality Experts Group (VQEG),* Tech. Rep., Mar. 2000.

[25] Z. Wang, Z. Wang, and A. C. Bovik, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2002, pp. 3313–3316.

[26] P. Longhurst and A. Chalmers, "User validation of image quality assessment algorithms," in *Proc. Theory and Practice of Computer Graphics (TPCG'04)*. Washington, DC: IEEE Comput. Soc., 2004, pp. 196–202.

[27] J. Lubin, *A Visual Discrimination Model for Imaging System Design and Evaluation*, E. Peli, Ed. Singapore: World Scientific, 1995.

[28] B. Li, G. W. Meyer, and R. V. Klassen, "A comparison of two image quality models," in *Proc. SPIE Human Vision and Electronic Imaging III,* 1998, vol. 3299, pp. 98–109.

[29] M. R. Bolin and G. W. Meyer, "A perceptually based adaptive sampling algorithm," in *Proc. 25th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'98)*. New York: ACM, 1998, pp. 299–309.

[30] B. Watson, A. Friedman, and A. McGaffey, "Measuring and predicting visual fidelity," in *Proc. 28th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'01)*. New York: ACM, 2001, pp. 213–220.

[31] M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg, "A perceptually based physical error metric for realistic image synthesis," in *Proc. 26th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'99)*. New York: ACM/Addison-Wesley, 1999, pp. 73–82.

[32] G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala, "Visual equivalence: Towards a new standard for image fidelity," in *ACM SIGGRAPH 2007 Papers (SIGGRAPH'07)*. New York: ACM, 2007, pp. 76-1–76-11.

[33] J. A. Ferwerda, P. Shirley, S. N. Pattanaik, and D. P. Greenberg, "A model of visual masking for computer graphics," in *Proc. 24th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH'97)*. New York: ACM /Addison-Wesley, 1997, pp. 143–152.

[34] S. Albin, G. Rougeron, B. Peroche, and A. Tremeau, "Quality image metrics for synthetic images based on perceptual color differences," *IEEE Trans. Image Processing*, vol. 11, no. 9, pp. 961–971, Sept. 2002.

[35] B. E. Rogowitz and H. E. Rushmeier, "Are image quality metrics adequate to evaluate the quality of geometric objects," in *Proc. SPIE Human Vision and Electronic Imaging VI*, San Jose, CA, 2001, pp. 340–348.

[36] S. Silva, B. S. Santos, J. Madeira, and C. Ferreira, "Perceived quality assessment of polygonal meshes using observer studies: A new extended protocol," in *Proc. SPIE Human Vision and Electronic Imaging XIII*, 2008, vol. 6806.

[37] I. Howard and B. Rogers, *Seeing in Depth*. New York: Oxford Univ. Press, 2008.

[38] H. E. Rushmeier, B. E. Rogowitz, and C. Piatko, "Perceptual issues in substituting texture for geometry," in *Proc. SPIE Human Vision and Electronic Imaging V*, B. E. Rogowitz and T. N. Pappas, Eds., 2000, vol. 3959, no. 1, pp. 372–383.

[39] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," in *Proc. IS&T/SPIE Conf. Human Vision and Electronic Imaging III*, SPIE, 1998, vol. 3299, pp. 180–191.

[40] ITU Recommendation bt.500-10, "Methodology for subjective assessment of the quality of television pictures," *ITU, ITU Recommendation bt.500-10*, 2000.

[41] ITU Recommendation p.910, "Subjective video quality assessment methods for multimedia applications," *ITU, ITU Recommendation p.910*, 1996.

[42] I. Cleju and D. Saupe, "Evaluation of supra-threshold perceptual metrics for 3d models," in *Proc. 3rd Symp. Applied Perception in Graphics and Visualization (APGV'06)*. New York: ACM, 2006, pp. 41–44.

[43] C. O'Sullivan, J. Dingliana, T. Giang, and M. K. Kaiser, "Evaluating the visual fidelity of physically based animations," *ACM Trans. Graph.*, vol. 22, pp. 527–536, July 2003.

[44] K. Myszkowski, "Perception-based global illumination, rendering, and animation techniques," in *Proc. 18th Spring Conf. Computer Graphics (SCCG'02)*. New York: ACM, 2002, pp. 13–24.

[45] A. Bulbul, C. Koca, T. Capin, and U. Güdükbay, "Saliency for animated meshes with material properties," in *Proc. 7th Symp. Applied Perception in Graphics and Visualization (APGV'10)*. New York: ACM, 2010, pp. 81–88.

[46] MyMultimediaWorld. [Online]. Available: http://www.mymultimediaworld.com

[47] M. Garland. (2009). QSlim simplification software. [Online]. Available: http://mgarland.org/software/qslim.html

[48] P. Cignoni et al. (2011). MeshLab. Visual Computing Lab, ISTI, CNR. [Online]. Available: http://meshlab.sourceforge.net/

[49] K. Wang, G. Lavoué, F. Denis, A. Baskurt, and X. He, "A benchmark for 3D mesh watermarking," in *Proc. IEEE Int. Conf. Shape Modeling and Applications,* 2010, pp. 231–235.

[50] S. Silva. (2008). PolyMeCo. [Online]. Available: http://www.ieeta.pt/polymeco/index.php/home

[51] AIM@SHAPE. [Online]. Available: http://www.aimatshape.net/

Paul Coverdale, Sebastian Möller, Alexander Raake, and Akira Takahashi

Multimedia Quality Assessment Standards in ITU-T SG12

[Models predicting quality based on parameters and bit stream information]

© INGRAM PUBLISHING

New media coding technologies enabling high-quality audio and video, network-based services, and the proliferation of low-cost end-point devices have made packet-based audiovisual services, such as mobile television (TV) and Internet Protocol TV (IPTV), increasingly popular. For service and network providers, it is important to be able to quantify the quality of experience (QoE) of an audiovisual service as perceived by the end user. Service planning, implementation, and monitoring should be based on the concept of QoE as an indicator of user satisfaction and high acceptance of the service, and not on the basis of quality of service (QoS), i.e., on network performance indicators. Whereas the latter are important prerequisites for high QoE, they are not necessarily directly linked to user perception and satisfaction.

Definitive quality assessment requires subjective testing but the only practical solution during service operation is to use an objective quality assessment model, which produces an estimate of the perceived quality in a measurement scenario. Whatever method is used, it is important that it is based on a well-accepted standard, so that results can be compared on an even basis. Study Group 12 of the Telecommunication Standardization Section of the International Tele-communication Union (ITU-T SG12) has been involved for many years in standardizing

Digital Object Identifier 10.1109/MSP.2011.942467
Date of publication: 1 November 2011

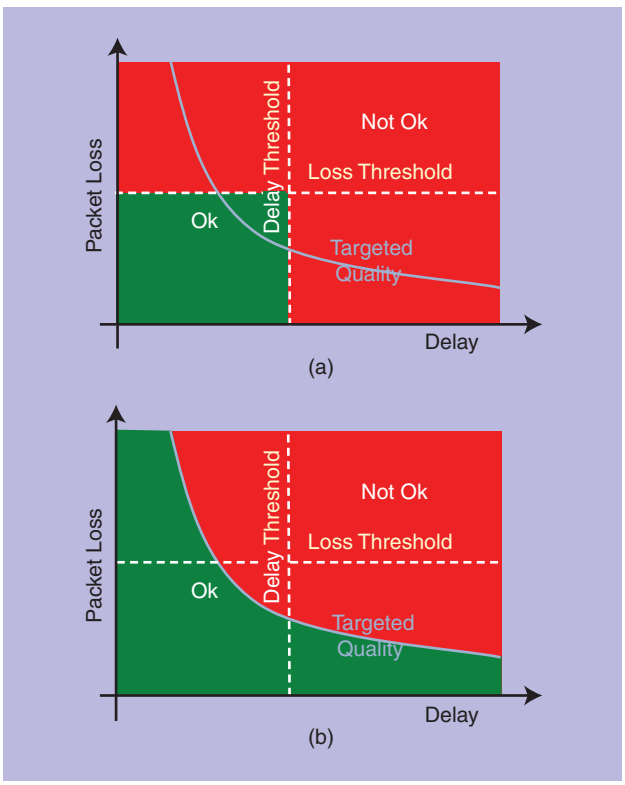
1053-5888/11/$26.00©2011IEEE

methods for multimedia quality assessment, both subjective and objective, with the goal to meet the needs and requirements of multimedia service providers and equipment vendors. This article gives an overview of existing and emerging SG12 standards in this area, with a special focus on models that predict quality on the basis of parameters and bit stream information that is available during network planning and monitoring phases.

E-MODEL QUALITY PREDICTIONS COVER THE COMBINED EFFECTS OF DIFFERENT DEGRADATIONS OCCURRING SIMULTANEOUSLY ON THE TELEPHONE LINE AND AT THE SENDING AND RECEIVING TERMINAL.

SPEECH QUALITY PLANNING MODELS

Speech transmission planning models have been in widespread use by network operators since the 1970s. At that time, each operator defined its own model, making it impossible to compare predictions across networks [14]. This situation was intolerable when networks were widely interconnected. In the early 1990s, the European Telecommunications Standards Institute (ETSI) Business TeleCommunications (BTC)2 made an effort to harmonize the predictions and to come up with a single model [15], which was later adopted and improved by ITU-T SG12, and is now widely known as the E-model (Recommendation G.107 [1]).



[FIG1] Schematic representation of the usefulness of a parametric quality prediction model to avoid the overengineering of networks, here for predicting the joint effects of packet loss and delay. (a) Planning based on individual threshold values for packet loss and delay. (b) Planning based on a parametric model such as the E-model. Green areas are the ones considered to be "acceptable" with each approach; the curved line shows the actually limiting targeted quality threshold (inspired by [21]).

The E-model provides predictions for the overall quality experienced in a conversational situation on the basis of parameters that are estimated at the network planning stage. Input parameters consist of loudness ratings (i.e., frequency-weighted average amplitude spectra for the transmission paths under consideration), average delay times (ignoring the phase distortions), as well as frequency-weighted noise levels (roughly reflecting the disturbance related to the respective noise). The output value is an estimation of the overall conversational quality, in terms of a transmission rating R varying between zero for the lowest possible quality to 100 for the optimum quality of a narrow-band (300–3,400 Hz) telephone connection. A standard integrated services digital network (ISDN) connection with a low noise floor and logarithmic pulse-code modulation (PCM) coding achieves a value of R = 93.2 on this scale, as can be calculated from the default parameter values given in Recommendation G.107 [1]. R values can be transformed to mean opinion scores (MOS-CQE for the estimated quality in a conversational situation) as they would be obtained in a standard conversation quality test according to [9] and [10], using a standard transformation law defined in [1]. Further transformations for R are defined for the expected percentage of users rating a connection good or better (GoB) or the percentage rating it poor or worse (PoW). R values can also be associated with speech transmission quality categories, as defined in Recommendation G.109 [2].

E-model quality predictions cover the combined effects of different degradations occurring simultaneously on the telephone line and at the sending and receiving terminal. As a result, the predictions are a much better estimation of the actual impact than single parameter value thresholds would be. In fact, an overengineering margin would occur if individual thresholds were defined for each parameter, which can be avoided with the use of parametric models such as the E-model. Figure 1 depicts this situation for the combination of packet loss and delay impairments, which is a common compromise to be made in planning IP-based networks: The longer the jitter buffer and the corresponding delay, the lower the packet loss rate, but the higher the conversational impact due to delay. When defining fixed thresholds for each degradation individually, these thresholds would lead to the green "allowed area" in Figure 1(a). However, the perceptually relevant threshold when both degradations occur simultaneously would be the curved line; this would allow even higher packet loss rates when delay is low, or higher delays when packet loss is low. Planning according to the individual thresholds for delay and packet loss would violate the perceptual threshold in some cases, and it would be too restrictive in other cases. This

can be avoided by directly taking the perceptual threshold (resulting from the combination of degradations) as the reference, as is done in the E-model.

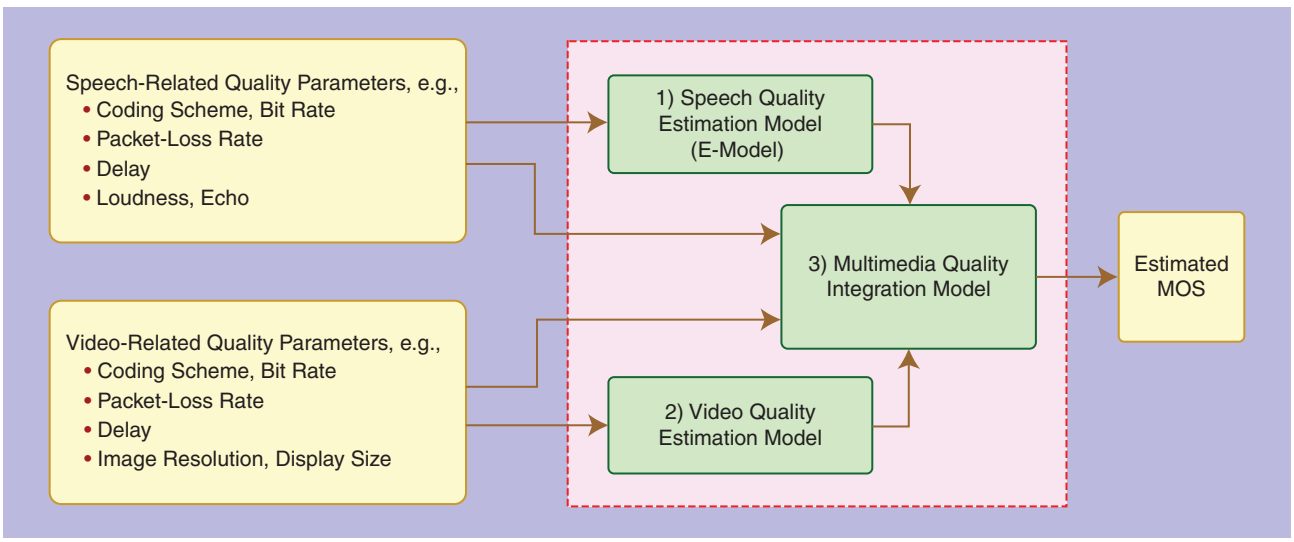
In its current state, the E-model considers degradations due to nonoptimum loudness, noise, nonlinear codec distortions, echo, absolute delay, sidetone, and packet loss in IP-based networks. Each type of degradation is transformed to a so-called psychological scale (the transmission rating scale) where it is expected to be additive to other types of degradations; on this scale, it determines a so-called "impairment factor," which quantifies the degradation in terms of *R* value units associated with that particular degradation. The overall transmission rating is then determined by subtracting all impairment factors from the maximum possible quality limited by the signal-to-noise ratio of the connection. This principle is called the "impairment factor principle" and is described in more detail in [3].

Whereas most input parameters to the E-model can be determined through instrumental measurements—such as loudness ratings using the procedure described in [8], noise levels using a psophometer defined in [7]—one of the most important degradation occurring in today's digital networks resisted such an instrumental measurement until recently: nonlinear coding degradations, potentially in conjunction with packet loss (concealed or not), could only be judged in subjective tests, and results of these tests were necessary to derive input parameters to the E-model for such degradations (see Recommendation P.833 [19]). Today, the ITU-T recommends signal-based models for speech quality, using the input and output signals of the piece of device under test, specifically the codec, to estimate the inherent quality degradation, without any further human intervention.

THE CONVERSATIONAL IMPACT OF ABSOLUTE DELAY IS A PARTICULARLY DIFFICULT TOPIC, SINCE DELAY EFFECTS ARE DIFFICULT TO QUANTIFY AND FREQUENTLY OVERESTIMATED BY APPLYING STANDARD LIMITS.

The recommended signal-based model for this purpose is the Perceptual Evaluation of Speech Quality (PESQ), standardized in [12]. This standard has recently been augmented by a new standard [13]. Based on results from these models, an impairment factor for the codec (so-called equipment impairment factor, *Ie*) can be estimated in a purely instrumental way, using the normalization procedure given in [11]. *Ie* values can be further modified to cater for the effects of random or bursty packet loss, using the parameter *Bpl* (packet-loss robustness). *Bpl* values can also be determined in an instrumental way, using the procedure of [11]. Thus, all input values to the E-model are finally based on instrumental measurements, mostly intrusive ones to be performed before a network has been set up.

Whereas the E-model has originally been developed for 3.1 kHz handset telephony, first steps have been made in ITU-T SG12 to extend it to wideband (50–7,000 Hz) transmission scenarios. This includes the determination of a new maximum possible quality value on the *R* scale, which is currently set to 129 for the clean wideband channel. In other words, a clean wideband connection is expected to deliver a quality level which is around 30% higher than the one of a clean narrow-band connection. Initial experimental results suggest that this level can further be increased by going toward super-wideband (50–14,000 Hz) transmission, but these results are only preliminary at this stage [18]. Furthermore, the E-model has been updated for the effects of noise, linear degradations, nonlinear codec degradations, and packet loss in wideband transmission; see [17]. Current work includes addressing the effects of echoes and absolute delay, as well as the effects associated with terminal equipment



[FIG2] Block diagram of G.1070 model.

other than handsets, including the signal-processing equipment integrated into such equipment (noise reduction, echo cancellation); see [16]. The conversational impact of absolute delay is a particularly difficult topic, since delay effects are difficult to quantify and frequently overestimated by applying standard limits [4].

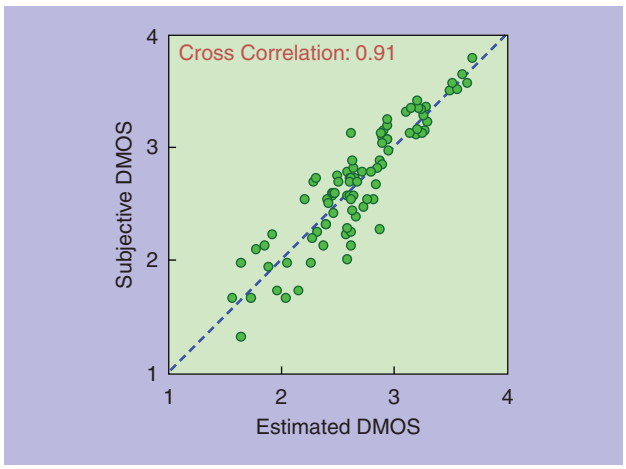
IT IS IMPORTANT THAT QUALITY ASSESSMENT IS BASED ON A WELL-ACCEPTED STANDARD, SO THAT RESULTS CAN BE COMPARED ON AN EVEN BASIS.

VIDEO QUALITY PLANNING MODELS

Recommendation G.1070 [6] defines an opinion model for network planning and terminal design of video-phone applications. This is a counterpart of Recommendation G.10 for conventional telephony. As shown in Figure 2, the G.1070 model consists of three building blocks, which are speech, video, and multimedia quality estimation modules.

The speech quality estimation block is identical to the E-model [1] except for the function related to delay impairment, which is taken into account in the multimedia quality estimation module in G.1070. As for video compression impairment, G.1070 Annex A provides a means for determining the parameter values used in the model for specific video resolutions and video codecs, because the video quality heavily depends on individual codec implementations. In addition, G.1070 provides some examples of the parameter values in its Appendix (MPEG-4 in Quarter Video Graphics Array (QVGA) and Quarter QVGA (QQVGA) and MPEG-2 in VGA).

The quality estimation accuracy of the G.1070 model is demonstrated in Figure 3, based on experimental results reported in [20]. This data set was not used in the optimization of the model, which means that it is unknown to the model. The test conditions include not only audio/video coding, but also packet loss in IP networks, and delay including audio-visual asynchrony. The model shows a high correlation with the actual subjective judgment represented by subjective degradation MOS (DMOS).



[FIG3] Quality estimation accuracy of G.1070 model (used with permission from [20]).

ITU-T SG12 is currently studying such an opinion model for video streaming applications including IPTV, which is provisionally called G.OMVAS, where OMVAS stands for "opinion model for video and audio streaming" applications. It is expected that this model will be different from the G.1070 model because of the following three reasons:

1) Video quality quite heavily depends on the video signal to be transmitted; thus, relatively stable head-and-shoulder videos expected in a video-telephony situation will be less vulnerable to quality degradations than, e.g., highly dynamic movie sequences.
2) The sound associated with video streaming applications is not limited to speech alone; as a consequence, different types of codecs will be used, which are not foreseen by the E-model underlying the audio part of the G.1070 model.
3) Streaming applications are usually less critical to transmission delay than interactive communication applications.

AUDIO/VIDEO REAL-TIME QUALITY MONITORING MODELS

With its mix of participants from telecommunications service providers, equipment vendors, and test tool suppliers, ITU-T SG12 is in a good position to develop audio/video real-time quality monitoring models. There are currently work items in progress on two new recommendations for objective quality assessment models. The first work item, called P.NAMS, (where NAMS stands for nonintrusive parametric model for the assessment of performance of multimedia streaming) targets the standardization of a parametric, transport layer, nonintrusive audiovisual quality model. P.NAMS has two main application areas: mobile applications (including mobile TV and mobile video streaming) and wireline IPTV applications. The P.NAMS model will produce three quality estimation scores: audio, video, and audiovisual scores, provided on the five-point MOS scale. As in G.1070, the model will contain three basic building blocks for video, audio, and audiovisual quality estimation.

The P.NAMS model takes transport header information, prior knowledge about the services, and some side information from the client as input. The model will not be able to fully estimate the quality variation due to content variation, and implicitly a parametric model will not give the same quality prediction performance as a signal-based model for a large variety of conditions. On the other hand, the P.NAMS model is expected to require little computational resources, which makes implementation in many locations within the network or in an end-point device possible. The model can be deployed in situations where currently no audiovisual quality estimation is possible, enabling QoE assessment with acceptable prediction performance, potentially for a very large number of concurrent audiovisual sessions

in a network.The P.NAMS model can be deployed in four different ways as follows:

- The model is located inside the network along the media delivery path, and measurements are done at the connection point of the delivery path.
- The model is located inside the network along the media delivery path, and measurements are done at the connection point and at the end point. The end-point measurements are sent to the model with a standardized protocol.
- The model is located inside the network along the media delivery path, and measurements are done only at the end point and sent to the model with a standardized protocol.
- The model is located at the end point and has access to all the measurement data from the end point.

These situations are shown in Figure 4(a)–(d).

The four modes of operation cover a wide range of deployment scenarios, and enable implementation in a wide range of network products. The P.NAMS model is also intended to handle scenarios where error correction techniques, such as forward error correction (FEC), are used. The error-corrected stream is either reconstructed and used as input to the model, or the input parameters to the model are mapped to reflect the parameter values after error correction. The P.NAMS standard is expected to be finalized in 2012 as the result of a competitive model selection process.

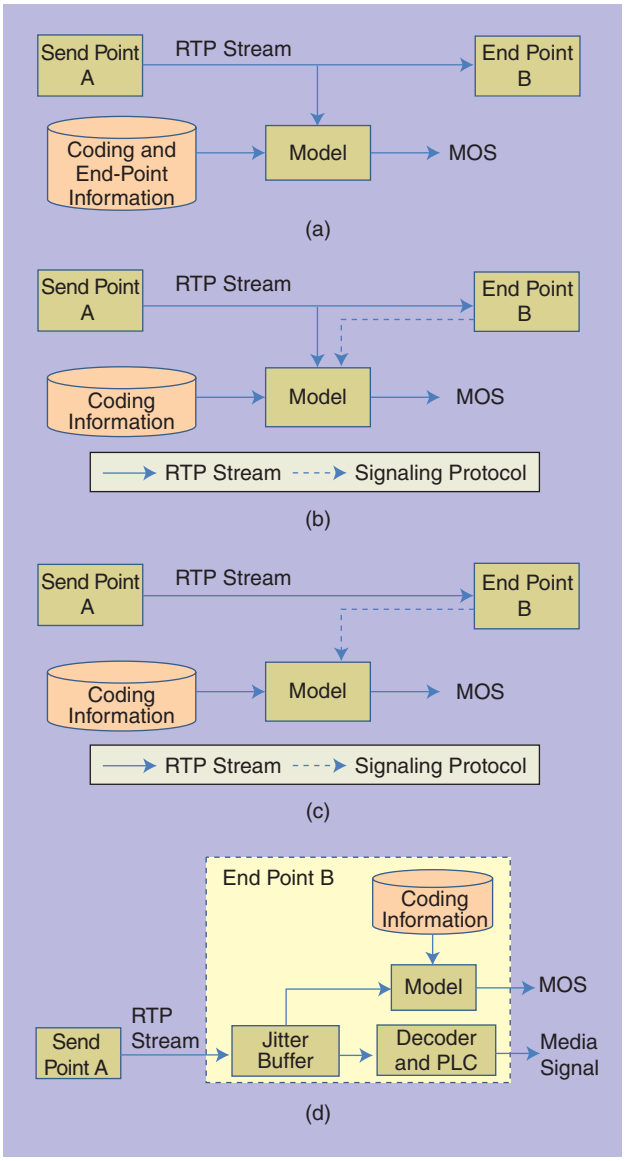
The second work item in Question 14 is a bit stream-layer video quality model called P.NBAMS (where NBAMS stands for nonintrusive bit stream model for the assessment of performance of multimedia streaming). It uses all information available to the packet-layer P.NAMS model, but in addition analyzes the payload (bit stream) of the video packets. This is assumed to enable a higher prediction performance than P.NAMS, but also yields slightly higher computational requirements. By analyzing the bit stream, it is possible to detect quality variations due to video content and encoding efficiency, as well as to obtain more detailed information about the effects due to packet loss. The P.NBAMS model has two modes: Mode 1 is a parsing-only, lower complexity mode not using pixel-level information, while the computationally more complex Mode 2 may involve partial or full decoding of video frames. It is expected that Mode 2 will have slightly better prediction performance than Mode 1.

The P.NBAMS model can potentially be deployed using the same four deployment settings as P.NAMS. In addition, it is expected that the P.NBAMS video-only quality model can be combined with the audio and audiovisual building blocks of P.NAMS to enable parametric audiovisual and audio quality estimation combined with a higher performing video quality model. A similar approach as used for P.NAMS for including FEC impact on the quality will most likely be used also for P.NBAMS. The P.NBAMS standard is expected to be ready in 2012, as a result of a joint competitive model selection for both Mode 1 and Mode 2 models.

ATTENTION IN SG12 IS NOW HEAVILY FOCUSED ON LIGHTWEIGHT MODELS BASED ON PARAMETRIC OR BIT STREAM INPUT.

QoE ASSESSMENT METHODOLOGIES

QoE is a widely used term in the telecommunications industry. However, until recently there has been no single document available that describes methods for assessing QoE. Recommendation G.1011 [5] provides such a reference guide to existing standardized methods in ITU-T for QoE assessment. The overall approaches for the different assessment methodologies typically used are described, and a taxonomy of QoE assessment standards is defined. For different applications, the recommendation identifies suitable subjective assessment methodologies, and where applicable, objective



[FIG4] (a) Network model-network measurements, (b) network model-network/end-point measurements, (c) network model-end-point measurements, and (d) end-point model-end-point measurements.

assessment methods that can be used to estimate subjective opinion, and gives guidance on their usage and limitations. The intended users of this recommendation include carriers, service providers, and equipment manufacturers. Figure 5 provides a summary of how existing ITU recommendations can be applied for QoE assessment, making use of both subjective and objective testing methodologies, covering a range of applications; media type; interactivity (conversational (CONV)/nonconversational (NONCONV), important for appropriately taking into account the effects of delay); reference signal availability [full reference (FR), reduced reference (RR), no reference (NR)]; and usage [laboratory (LAB), monitoring (MON), and planning (PLN)].

THESE MODELS HAVE A POTENTIAL TO SIMULTANEOUSLY MONITOR MANY, OR EVEN ALL, OF THE AUDIOVISUAL SESSIONS IN A NETWORK.

SUMMARY

As multimedia services become more widespread, there is a need for standardized methods for assessment of the quality perceived by the end user. ITU-T SG12 has been at the forefront in defining and validating subjective and objective quality assessment methods for many years and has issued many recommendations in this area. Subjective assessment remains the definitive methodology but objective methods are now becoming mature enough to provide very accurate results, especially for voice, and much more convenient to implement. However, currently available objective quality estimation methods can require a lot of computational resources, which makes them difficult, or even impossible, to be deployed operationally in a large network. Attention in SG12 is now heavily focused on lightweight models based on parametric or bit stream input, which enable large-scale real-time QoE assessment in a network. These models have a potential to simultaneously monitor many, or even all, of the audiovisual sessions in a network.

Application	Media	Conversational (CONV)/Non-conversational (NONCONV)	Subjective Test Methodology	Objective Test Methodology		
				Model	FR/RR/NR	Primary Usage
Telephony	Speech	NONCONV	[ITU-T P.800] [ITU-T P.830] [ITU-T P.835]	[ITU-T P.862] + [ITU-T P.862.1] (NB) [ITU-T P.862.2] (WB)	FR	LAB, MON
				[ITU-T P.563] (NB) [ITU-T P.564] (NB/WB)	NR	MON
		CONV	[ITU-T P.800] [ITU-T P.805]	[ITU-T G.107] (NB)	NR	PLN
				[ITU-T P.561] + [ITU-T P.562] (NB/WB)	NR	MON
Video-Telephony	Multimedia (Note)	CONV	[ITU-T P.920]	[ITU-T G.1070] (NB/WB)	NR	PLN
Video-Streaming (Mobile TV/IPTV)	Video	NONCONV	[ITU-T P.910] [ITU-T J.140] [ITU-R BT.500-12]	[ITU-T J.144] (SDTV) [ITU-T J.247] (QCIF, CIF, VGA)	FR	LAB, MON
				[ITU-T J.249] (SDTV) [ITU-T J.246] (QCIF, CIF, VGA)	RR	MON
	Audio	NONCONV	[ITU-T P.830] [ITU-R BS.1116-1] [ITU-R BS.1285] [ITU-R BS.1534-1]	[ITU-R BS.1387]	FR/RR	MON/PLN
	Multimedia	NONCONV	[ITU-T P.911]			
Web Browsing	Data			[ITU-T G.1030]	NR	PLN

NOTE: For individual media (i.e., speech and video), the recommendations used in telephony and video-streaming applications are applicable.

[FIG5] **Current ITU recommendations for QoE assessment (data taken from Table 9-1 of [5]).**

AUTHORS

Paul Coverdale (coverdale@sympatico.ca) received an honors B.Sc. degree in electrical and electronic engineering in 1971 from the University of Leeds, England. After spending 30 years with Nortel Networks, where he held a number of engineering and management positions related to the specification, design, verification and standardization of wireline and wireless products, he now acts as a consultant in the area of QoS/QoE standards. He has been contributing to ITU-T SG 12 since 1983, where he has held many leadership positions. He is currently chair of WP3/12 "Multimedia Qos and QoE."

Sebastian Möller (sebastian.moeller@telekom.de) studied electrical engineering at the universities of Bochum (Germany), Orléans (France), and Bologna (Italy). He received a doctor-of-engineering degree in 1999 and the venia legendi for a book on the quality of telephone-based spoken dialogue systems in 2004. In 2005, he joined Deutsche Telekom Laboratories, Technical University (TU) Berlin, and in 2007, he was appointed professor for quality and usability at TU Berlin. His primary interests are in speech signal processing, speech technology, and quality and usability evaluation. Since 1997, he has taken part in ITU-T SG 12, where he is currently corapporteur of Q.8/12 "E-Model extension toward wideband transmission and future telecommunication and application scenarios."

Alexander Raake (alexander.raake@telekom.de) received his doctoral degree in electrical engineering and information technology from Ruhr-University Bochum, Germany, in 2005, and his electrical engineering diploma from RWTH Aachen, Germany, in 1997. From 1998 to 1999 he was a researcher at EPFL, Switzerland. Between 2004 and 2009 he held postdoc and senior scientist positions at LIMSI-CNRS, France, and Deutsche Telekom Laboratories, Germany, respectively. Since 2009, he has been an assistant professor at Deutsche Telekom Laboratories, TU Berlin. His research interests are in multimedia technology and QoE. He has been active in ITU-T since 1999; currently he is corapporteur of Q.14/12 "Development of Parametric Models and Tools for Audiovisual and Multimedia Quality Measurement Purposes."

Akira Takahashi (takahashi.akira@lab.ntt.co.jp) received the B.S. degree in mathematics from Hokkaido University in Japan in 1988, M.S. degree in electrical engineering from the California Institute of Technology in the United States in 1993, and Ph.D degree in engineering from the University of Tsukuba in Japan in 2007. He joined NTT Laboratories in 1988 and has been engaged in the quality assessment of audio and visual communications. Currently, he is the manager of the Service Assessment Group in NTT Service Integration Laboratories. He has contributed to ITU-T SG12 since 1994. He is currently a vice chair of ITU-T SG12, and rapporteur for Q.13/12 "QoE, QoS, and Performance Requirements and Assessment Methods for Multimedia Including IPTV."

REFERENCES

[1] ITU, "The E-model: A computational model for use in transmission planning," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.107, Apr. 2009.

[2] ITU, "Definition of categories of speech transmission quality," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.109, Sept. 1999.

[3] ITU, "Transmission impairments due to speech processing," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.113, Sept. 2007.

[4] ITU, "One-way transmission time," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.114, May 2003.

[5] ITU, "Reference guide to quality of experience assessment methodologies," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.1011, June 2010.

[6] ITU, "Opinion model for video-telephony applications," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. G.1070, Apr. 2007.

[7] ITU, "Psophometer for use on telephone-type circuits," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. O.41, Oct. 1994.

[8] ITU, "Determination of loudness ratings; fundamental principles," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.76, Nov. 1988.

[9] ITU, "Methods for subjective determination of transmission quality," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.800, Aug. 1996.

[10] ITU, "Subjective evaluation of conversational quality," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.805, Apr. 2007.

[11] ITU, "Methodology for the derivation of equipment impairment factors from instrumental models," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.834, July 2002.

[12] ITU, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.862, Feb. 2011.

[13] ITU, "Perceptual objective listening quality assessment," Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.863, Jan. 2011.

[14] ITU, "Models for predicting transmission quality from objective measurements," Int. Telecomm. Union, Geneva, Switzerland, ITU–T Suppl. 3 to P–Series Rec, 1993 (superseded).

[15] N. O. Johannesson, "The ETSI computation model: A tool for transmission planning of telephone networks," *IEEE Commun. Mag.*, vol. 35, no. 1, pp. 70–79, Jan. 1997.

[16] S. Möller, F. Kettler, H.-W. Gierlich, N. Côté, A. Raake, and M. Wältermann, "Extending the E-model to better capture terminal effects," in *Proc. 3rd Int. Workshop Perceptual Quality of Systems (PQS'10)*, Bautzen, Germany, Sept. 6–8, 2010.

[17] A. Raake, S. Möller, M. Wältermann, N. Côté, and J.-P. Ramirez, "Parameter-based prediction of speech quality in listening context–Towards a WB E-model, in *Proc. 2nd Int. Workshop Quality of Multimedia Experience (QoMEX'10)*, Trondheim, Norway, June 21–23, 2010, pp. 182–187.

[18] M. Wältermann, I. Tucker, A. Raake, and S. Möller, "Extension of the E-model towards super-wideband speech transmission," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'10)*, Dallas, TX, Mar. 14–19, 2010, pp. 4654–4657.

[19] ITU, "Methodology for derivation of equipment impairment factors from subjective listening-only tests," *Int. Telecomm. Union, Geneva, Switzerland, ITU-T Rec. P.833*, 2001–2002.

[20] T. Hayashi, K. Yamagishi, T. Tominaga, and A. Takahashi, "Multimedia quality integration function for videophone services," in *Proc. GLOBECOM,* 2007, pp. 2735–2739.

[21] V. Barriac, "Non-intrusive voice transmission quality measurement techniques," in *Proc. DAGA,* 2002.

[SP]

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

[Moshe Mishali and Yonina C. Eldar]

Sub-Nyquist Sampling

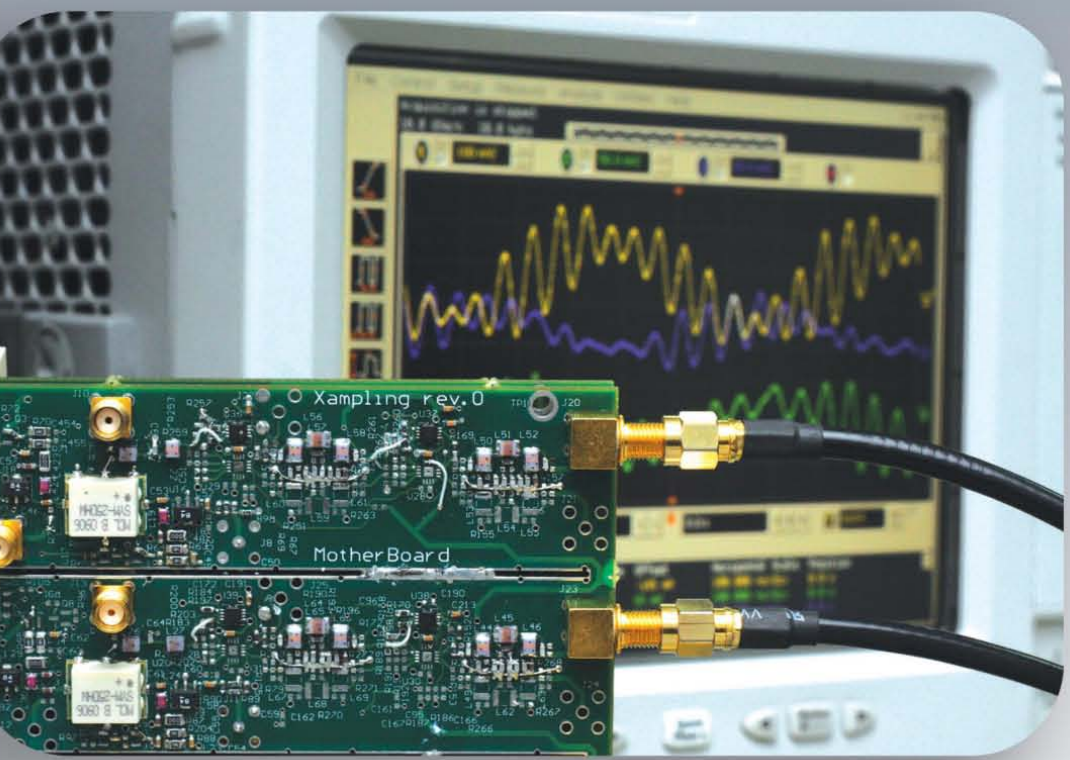


PHOTO COURTESY OF MOSHE MISHALI

[Bridging theory and practice]

Signal processing methods have changed substantially over the last several decades. In modern applications, an increasing number of functions is being pushed forward to sophisticated software algorithms, leaving only delicate finely tuned tasks for the circuit level. Sampling theory, the gate to the digital world, is the key enabling this revolution, encompassing all aspects related to the conversion of continuous-time signals to discrete streams of numbers. The famous Shannon-Nyquist theorem has become a landmark: a mathematical statement that has had one of the most profound impacts on industrial development of digital signal processing (DSP) systems.

Over the years, theory and practice in the field of sampling have developed in parallel routes. Contributions by many research groups suggest a multitude of methods, other than uniform sampling, to acquire analog signals [1]–[6]. The math has deepened, leading to abstract signal spaces and innovative sampling techniques. Within generalized sampling theory, bandlimited signals have no special preference, other than historic. At the same time, the market adhered to the Nyquist paradigm;

Digital Object Identifier 10.1109/MSP.2011.942308
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE

state-of-the-art analog to digital conversion (ADC) devices provide values of their input at equalispaced time points [7], [8]. The footprints of Shannon-Nyquist are evident whenever conversion to digital takes place in commercial applications.

Today, seven decades after Shannon published his landmark result in [9], we are witnessing the outset of an interesting trend. Advances in related fields, such as wideband communication and radio-frequency (RF) technology, open a considerable gap with ADC devices. Conversion speeds that are twice the signal's maximal frequency component have become more and more difficult to obtain. Consequently, alternatives to high rate sampling are drawing considerable attention in both academia and industry.

In this article, we review sampling strategies that target reduction of the ADC rate below Nyquist. Our survey covers classic works from the early 1950s through recent publications from the past several years. The prime focus is bridging theory and practice, that is, to pinpoint the potential of sub-Nyquist strategies to emerge from the math to the hardware. In this spirit, we integrate contemporary theoretical viewpoints, which study signal modeling in a union of subspaces, together with a taste of practical aspects, more specifically how the avant-garde modalities boil down to concrete signal processing systems. Our hope is that this presentation style will attract the interest of both researchers and engineers with the aim of promoting the sub-Nyquist premise into practical applications while encouraging further research into this exciting new frontier.

INTRODUCTION

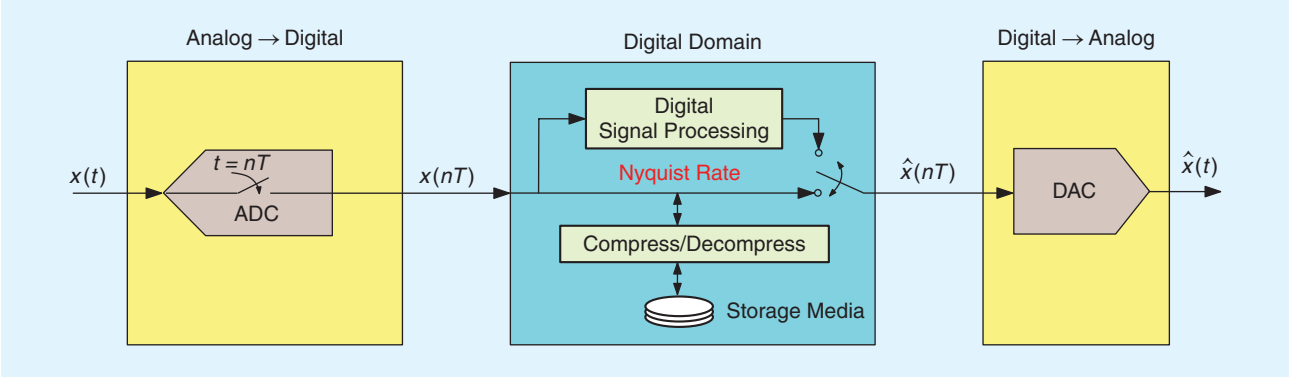
We live in a digital world. Telecommunication, entertainment, gadgets, and business all revolve around digital devices. These miniature sophisticated black boxes process streams of bits accurately at high speeds. Nowadays, it feels natural for electronic consumers when a media player shows their favorite movie, or when their surrounding system synthesizes pure acoustics, as if sitting in the orchestra and not in the living room. The digital world plays a fundamental role in our everyday routine, to such a point that we almost forget that we cannot "hear" or "watch" these streams of bits running behind the scenes. The world around us is analog, yet almost all modern man-made means for exchanging information are digital. "I am an analog girl in a digital world," sings Judy Gorman (One Sky Music International, 1998), capturing the essence of the digital revolution.

ADC technology lies at the heart of this revolution. ADC devices translate physical information into a stream of numbers, enabling digital processing by sophisticated software algorithms. The ADC task is inherently intricate: its hardware must hold a snapshot of a fast-varying input signal steady, while acquiring measurements. Since these measurements are spaced in time, the values between consecutive snapshots are lost. In general, therefore, there is no way to recover the analog input unless some prior on its structure is incorporated.

A common approach in engineering is to assume that the signal is bandlimited, meaning that the spectral contents are confined to a maximal frequency $f_{\max}$. Bandlimited signals have limited (hence slow) time variation, and can therefore be perfectly reconstructed from equalispaced samples with a rate at least $2f_{\max}$, termed the Nyquist rate. This fundamental result is often attributed in the engineering community to Shannon-Nyquist [9], [10], although it dates back to earlier works by Whittaker [11] and Kotel´nikov [12].

In a typical signal processing system, a Nyquist ADC device provides uniformly spaced pointwise samples $x(nT)$ of the analog input $x(t)$, as depicted in Figure 1. In the digital domain, the stream of numbers is either processed or stored. Compression is often used to reduce storage volume. DSP, which is unquestionably the crowning glory of this flow, is typically performed on the uncompressed stream. The delicate interaction with the continuous world is isolated to the ADC stage, so that sophisticated algorithms can be developed in a flexible software environment. The flow of Figure 1 ends with a digital to analog (DAC) device that reconstructs $x(t)$ from the high Nyquist-rate sequence $x(nT)$.

A fundamental reason for processing at the Nyquist rate is the clear relation between the spectrum of $x(t)$ and that of $x(nT)$, so that digital operations can be easily substituted for their continuous counterparts. Digital filtering is an example where this relation is successfully exploited. Since the power spectral densities of continuous and discrete random



[FIG1] Conventional blocks in a DSP system.

processes are associated in a similar manner, estimation and detection of parameters of analog signals can be performed by DSP. In contrast, compression, in general, results in a nonlinear complicated relationship between $x(t)$ and the stored data.

THE BANDLIMITED SIGNAL MODEL IS A NATURAL CHOICE TO DESCRIBE PHYSICAL PROPERTIES THAT ARE ENCOUNTERED IN MANY APPLICATIONS.

This article reviews alternatives to the scheme of Figure 1, whose common denominator is sampling at a rate below Nyquist. Research on sub-Nyquist sampling spans several decades and has been attracting renewed attention lately, since the growing interest in sampling in union of subspaces, finite rate of innovation (FRI) models, and compressed sensing techniques. Our goal in this survey is to provide an overview of various sub-Nyquist approaches. We focus this presentation on one-dimensional signals $x(t)$, with applications to wideband communication, channel identification, and spectrum analysis. Two-dimensional imaging applications are also briefly discussed.

Throughout, the theme is bridging theory and practice. Therefore, before detailing the specifics of various sub-Nyquist approaches, we first discuss the relation between theory and practice in a broader context. The example of uniform sampling, which without a doubt crossed that bridge, is used to list the essential ingredients of a sampling strategy so that it has the potential to step from math to actual hardware. Our subsequent presentation of sub-Nyquist strategies attempts to give a taste from both worlds—presenting the theoretical principles underlying each strategy and how they boil down to concrete and practical schemes. Where relevant, we shortly elaborate on practical considerations, e.g., hardware complexity and computational aspects.

ESSENTIAL INGREDIENTS OF A SAMPLING SYSTEM

NYQUIST SAMPLING

In 1949, Shannon formulated the following theorem for "a common knowledge in the communication art" [9, Th. 1]:

> If a function $f(t)$ contains no frequencies higher than W cycles-per-second, it is completely determined by giving its ordinates at a series of points spaced $1/2W$ seconds apart.

It is instructive to break this one-sentence formulation into three pieces. The theorem begins by defining an analog signal model—those functions $f(t)$ that do not contain frequencies above W Hz. Then, it describes the sampling stage, namely pointwise equalispaced samples. In between, and to some extent implicitly, the required rate for this strategy is stated: at least $2W$ samples per second.

The bandlimited signal model is a natural choice to describe physical properties that are encountered in many applications. For example, a physical communication medium often dictates the maximal frequency that can be reliably transferred. Thus, material, length, dielectric properties, shielding, and other electrical parameters define the maximal frequency W. Often, bandlimitedness is enforced by a lowpass filter with cutoff W, whose purpose is to reject thermal noise beyond frequencies of interest.

The implementation suggested by the Shannon-Nyquist theorem, equalispaced pointwise samples of the input, is essentially what industry has been persistently striving to achieve in ADC design. The sampling stage, per se, is insufficient; The digital stream of numbers needs to be tied together with a reconstruction algorithm. The famous interpolation formula

$$f(t) = \sum_n f\left(\frac{n}{2W}\right)\mathrm{sinc}(2Wt - n), \quad \mathrm{sinc}(\alpha) \triangleq \frac{\sin(\pi\alpha)}{\pi\alpha}, \quad (1)$$

which is described in the proof of [9], completes the picture by providing a concrete reconstruction method. Although (1) theoretically requires infinitely many samples to recover $f(t)$ exactly, in practice, truncating the series to a finite number of terms reproduces $f(t)$ quite accurately [13].

The theory ensures perfect reconstruction from samples at rate $2W$. A generalized sampling theorem by Papoulis allows to relax design constraints by replacing a single Nyquist-rate ADC by a filter bank of M branches, each sampled at rate $2W/M$ [14]. Another route to design simplification is oversampling, which is often used to replace the ideal brick wall filter by more flexible filter designs and to combat noise. Certain ADC designs, such as sigma-delta conversion, intentionally oversample the input signal, effectively trading sampling rate for higher quantization precision. Our wish list, therefore, includes a similar guideline for sub-Nyquist strategies: achieve the lowest rate possible in an ideal noiseless setting, and relax design constraints by oversampling and parallel architectures.

Further to what is stated in the theorem, we believe that two additional ingredients motivate the widespread use of the Shannon theorem. First, the interpolation formula (1) is robust to various noise sources: quantization round-off, series truncation and jitter effects [13]. The second appeal of this theorem lies in the ability to shift processing tasks from the analog to the digital domain. DSP is perhaps the major driving force that supports the wide popularity of Nyquist sampling. In sub-Nyquist sampling, the digital stream is, by definition, different from the Nyquist-rate sequence $x(nT)$. Therefore, the challenge of reducing sampling rate creates another obstacle—interfacing the samples with DSP algorithms that are traditionally designed to work with the high-rate sequence $x(nT)$, without necessitating interpolation of the Nyquist-rate samples. In other words, we would like to perform DSP at the low sampling rate as well.

Table 1 summarizes a wish list for a sub-Nyquist system, based on those properties observed in the Shannon theorem. A

sampling strategy satisfying most of these properties can, hopefully, find its way into practical applications.

ARCHITECTURE OF A SUB-NYQUIST SYSTEM

A high-level architecture of a sub-Nyquist system is depicted in Figure 2, following the spirit of the traditional block diagram of Figure 1. The ADC task is carried out by some hardware mechanism, which outputs a sequence $y[n]$ of measurements at a low rate. Since the sub-Nyquist samples $y[n]$ are, by definition, different from the uniform Nyquist sequence $x(nT)$ of Figure 1, a digital core may be needed to preprocess the raw data before DSP can take place. A prominent advantage over conventional Nyquist architectures is that the DSP operations are carried out at the low input rate. The digital core may also be needed to assist in reconstructing $x(t)$ from $y[n]$. Another advantage is that storage may not require a preceding compression stage; conceptually, the compression has already been performed by the sub-Nyquist sampling hardware.

An important point that we would like to emphasize is that strictly speaking, none of the methods we survey actually breach the Shannon-Nyquist theorem. Sub-Nyquist techniques leverage known signal structure, that goes beyond knowledge of the maximal frequency component. The key to developing interesting sub-Nyquist strategies is to rely on structure that is not too limiting and still allows for a broad class of signals on the one hand, while enabling sampling rate reduction on the other. One of the earlier examples demonstrating how signal structure can lead to rate reduction is sampling of multiband signals with known center frequencies, namely, signals that consists of several known frequency bands. We begin our review with this classic setting. We then discuss more recent paradigms that enable sampling rate reduction even when the band positions are unknown. As we show, this setting is a special case of a more general class of signal structures known as unions of subspaces, which includes a variety of interesting examples. After introducing this general model, we consider several sub-Nyquist techniques that exploit such signal structure in sophisticated ways.

[TABLE 1] SUB-NYQUIST SAMPLING: A WISH LIST.

INGREDIENT	REQUIREMENT
SIGNAL MODEL	ENCOUNTERED IN APPLICATIONS
SAMPLING RATE	APPROACH THE MINIMAL FOR THE MODEL AT HAND
IMPLEMENTATION	HARDWARE: LOW COST, SMALL NUMBER OF DEVICES SOFTWARE: LIGHT COMPUTATIONAL LOADS, FAST RUN TIME
ROBUSTNESS	REACT GRACEFULLY TO DESIGN IMPERFECTIONS LOW SENSITIVITY TO NOISE
PROCESSING	ALLOW VARIOUS DSP TASKS

CLASSIC SUB-NYQUIST METHODS

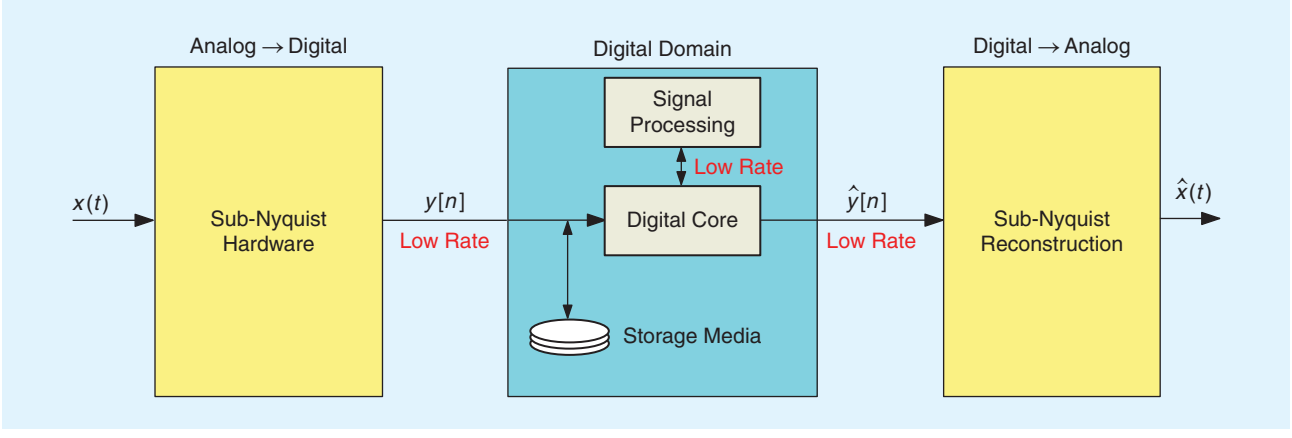
In this section, we survey classic sampling techniques which reduce the sampling rate below Nyquist, assuming a multiband signal with known frequency support. An example of a multiband input with N bands is depicted in Figure 3, with individual bandwidths not greater than B Hz, centered around carrier frequencies $f_i \le f_{\max}$ (N is even for real-valued inputs). Since the carriers f_i are known and the spectral support is fixed, the set of multiband inputs on that support is closed under linear combinations, thereby forming a subspace of possible inputs. Overlapping bands are permitted, though in practical scenarios, e.g., communication signals, the bands typically do not overlap.

DEMODULATION

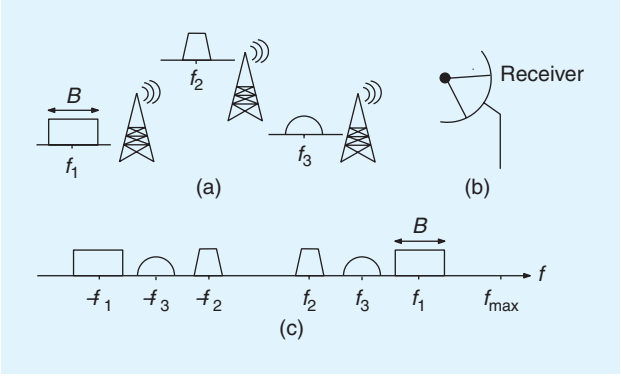
The most common practice to avoid sampling at the Nyquist rate,

$$f_{\text{NYQ}} = 2f_{\max}, \tag{2}$$

is demodulation. The signal $x(t)$ is multiplied by the carrier frequency f_i of a band of interest, so as to shift contents of a single narrowband transmission from high frequencies to the



[FIG2] **A high-level architecture of a sub-Nyquist system. Both processing and continuous recovery are based on lowrate computations. The raw data can be directly stored.**



[FIG3] **Parts (a)–(c) show three RF transmissions with different carriers f_i. In (c), the receiver sees a multiband signal.**

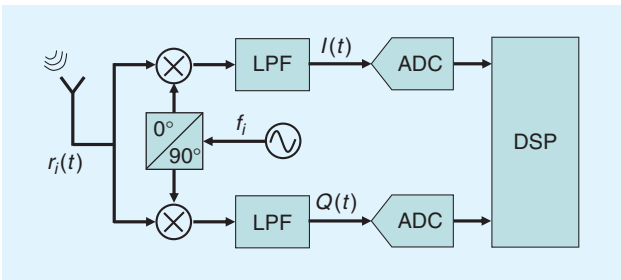
origin. This multiplication also creates a narrowband image around $2f_i$. Lowpass filtering is used to retain only the baseband version, which is subsequently sampled uniformly in time. This procedure is carried out for each band individually.

Demodulation provides the DSP block with the information encoded in a band of interest. To make this statement more precise, we recall how modern communication is often performed. Two $B/2$-bandlimited information signals $I(t), Q(t)$ are modulated on a carrier frequency f_i with a relative phase shift of 90°. The quadrature output signal is then given by [15]

$$r_i(t) = I(t)\cos(2\pi f_i t) + Q(t)\sin(2\pi f_i t). \tag{3}$$

For example, in amplitude modulation (AM), the information of interest is the amplitude of $I(t)$, while $Q(t) = 0$. Phase- and frequency-modulation (PM/FM) obey (3) such that the analog message is $g(t) = \arctan(I(t)/Q(t))$ [16]. In digital communication, e.g., phase- or frequency shift-keying (PSK/FSK), $I(t), Q(t)$ carry symbols. Each symbol encodes one, two, or more 0/1 bits. The I/Q-demodulator, depicted in Figure 4, basically reverts the actions performed at the transmitter side, which constructed $r_i(t)$. Once $I(t), Q(t)$ are obtained by the hardware, a pair of low-rate ADC devices acquire uniform samples at rate B. The subsequent DSP block can infer the analog message or decode the bits from the received symbols.

Reconstruction of each $r_i(t)$, and consequently recovery of the multiband input $x(t)$, is as simple as remodulating the



[FIG4] **A block diagram of a typical I-Q demodulator.**

information on their carrier frequencies f_i, according to (3). This option is used in relay stations or regenerative repeaters that decode the information $I(t), Q(t)$, use digital error correction algorithms, and then transform the signal back to high frequencies for the next transmission section [17].

I/Q demodulation has different names in the literature: zero-IF receiver, direct conversion, or homodyne; cf. [15] for various demodulation topologies. Each band of interest requires two hardware channels to extract the relevant $I(t), Q(t)$ signals. A similar principle is used in low-IF receivers, which demodulate a band of interest to low frequencies but not around the origin. Low-IF receivers require only one hardware channel per band, though the sampling rate is higher compared to zero-IF receivers.

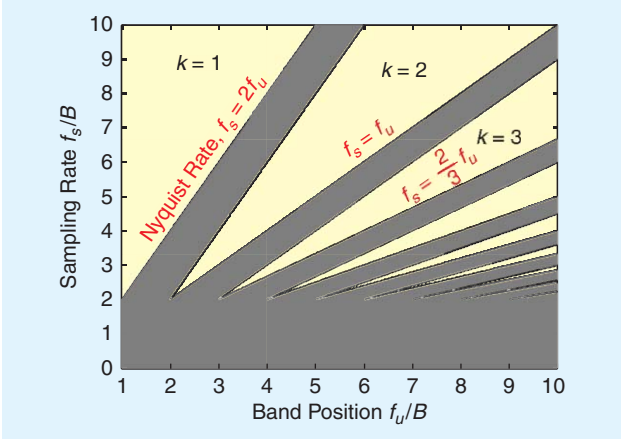
UNDERSAMPLING ADC

Aliasing is often considered an undesired effect of sampling. Indeed, when a bandlimited signal is sampled below its Nyquist rate, aliases of high-frequency content trample information located around other spectral locations and destroy the ability to recover the input. Undersampling (also known as direct bandpass sampling) refers to uniform sampling of a bandpass signal at a rate lower than the maximal frequency, in which case proper sampling rate selection renders aliasing advantageous.

Consider a bandpass input $x(t)$ whose information band lies in the frequency range (f_l, f_u) of length $B = f_u - f_l$. In this case, the lowest rate possible is $2B$ [19]. Uniform sampling of $x(t)$ at a rate of f_s that obeys

$$\frac{2f_u}{k} \le f_s \le \frac{2f_l}{k-1}, \tag{4}$$

for some integer $1 \le k \le f_u/B$, ensures that aliases of the positive and negative contents do not overlap [18]. Figure 5 illustrates the valid sampling rates implied by (4). In particular, the figure and (4) show that $f_s = 2B$ is achieved only if $x(t)$ has an integer band positioning, $f_u = kB$. Furthermore, as the rate reduction factor k increases, the valid region of



[FIG5] **The allowed (white) and forbidden (gray) undersampling rates of a bandpass signal depend on its spectral position [18] (figure courtesy of the IEEE.)**

sampling rates becomes narrower. For a given band position f_u, the region corresponding to the maximal $k \le f_u/B$ is the most sensitive to slight deviations in the exact values of f_s, f_l, f_u [18]. Consequently, besides the fact that $f_s = 2B$ cannot be achieved in general (even in ideal noiseless settings), a significantly higher rate is likely to be required in practice to cope with design imperfections.

Bridging theory and practice, the fact that (4) allows rate reduction, even though higher than the minimal, is useful in many applications. In undersampling, the ADC is applied directly to $x(t)$ with no preceding analog preprocessing components, in contrary to the RF hardware used in I/Q demodulation. However, not every ADC device fits an undersampling system: only those devices whose front-end analog bandwidth exceeds f_u are viable.

Undersampling has two prominent drawbacks. First, the resulting rate reduction is generally significantly higher than the minimal as evident from Figure 5. As listed in Table 1, approaching the minimal rate, at least theoretically, is a desired property. Second, and more importantly, undersampling is not suited to multiband inputs. In this scenario, each individual band defines a range of valid values for f_s according to (4). The sampling rate must be chosen in the intersection of these conditions. Moreover, it should also be verified that the aliases due to the different bands do not interfere. As noted in [21], satisfying all these constraints simultaneously, if possible, is likely to require a considerable rate increase.

PERIODIC NONUNIFORM SAMPLING

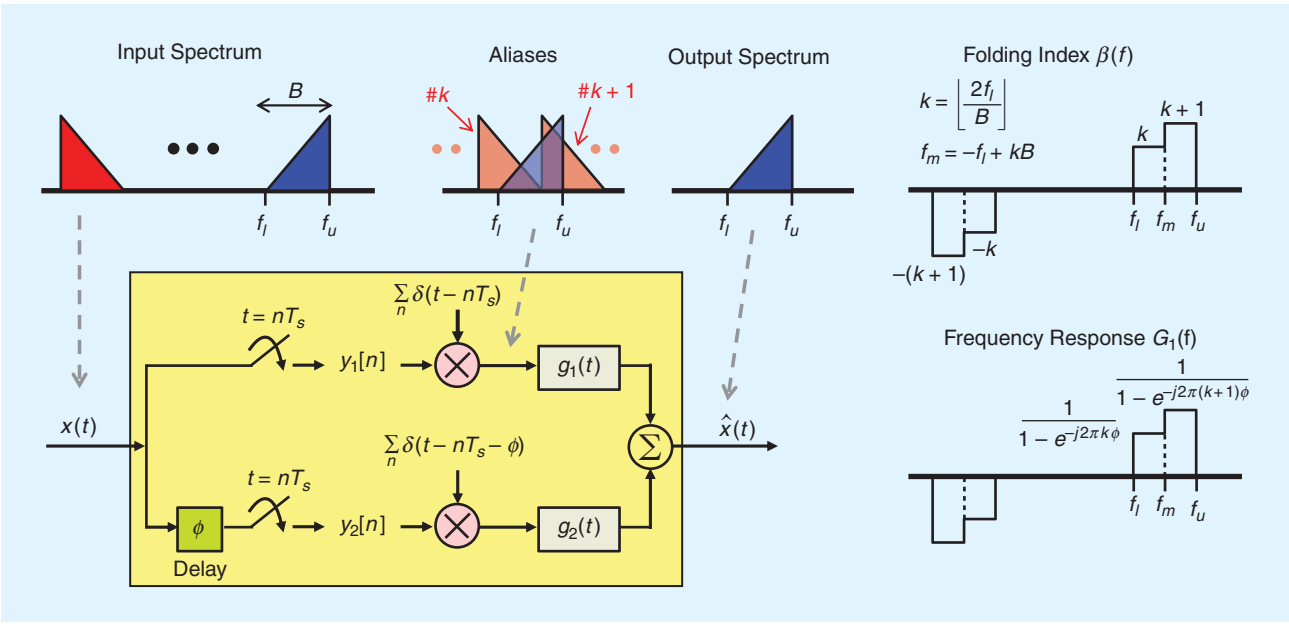
The discussion above suggests that uniform sampling may not be the most desirable acquisition strategy for inputs with multiband structure, unless sufficient analog hardware is used as in Figure 4. Classic studies in sampling theory have focused on nonuniform alternatives. In 1967, Landau proved a lower bound on the sampling rate required for spectrally sparse signals [19] with known frequency support when using pointwise sampling. In particular, Landau's theorem supports the intuitive expectation that a multiband signal $x(t)$ with N information bands of individual widths B necessitates a sampling rate no lower than the sum of the band widths, i.e., NB.

Periodic nonuniform sampling (PNS) allows to approach the minimal rate NB without complicated analog preprocessing. Besides ADC devices, the hardware needs only a set of time-delay elements. PNS consists of m undersampling sequences with relative time-shifts

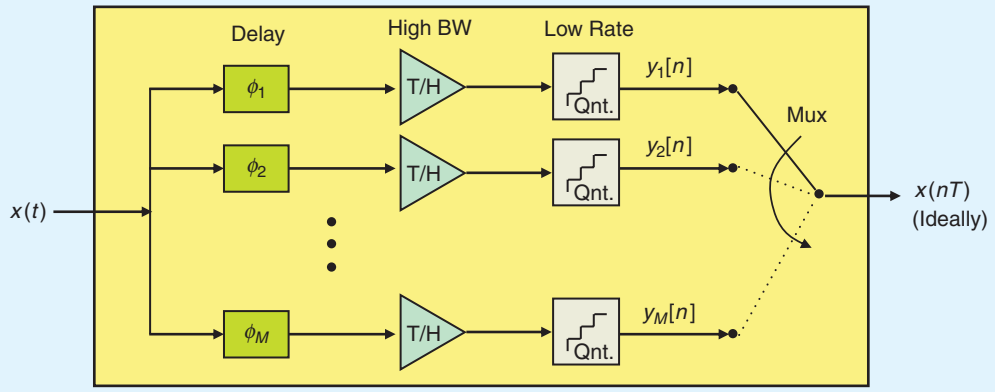
$$y_i[n] = x(nT_s + \phi_i), \quad 1 \le i \le m, \tag{5}$$

such that the total sampling rate m/T_s is lower than f_{NYQ}. Kohlenberg [22] was the first to prove perfect recovery of a bandpass signal from PNS samples taken at an average rate of $2B$ samples/s. Lin and Vaidyanathan [23] extended his approach to multiband signals.

We follow the presentation in [23] and explain how the parameters m, T_s, ϕ_i are chosen in the simpler case of a bandpass input. Suppose $x(t)$ is supported on $\mathcal{I} = (f_l, f_u) \cup (-f_u, -f_l)$ and $B = f_u - f_l$. We choose a PNS system with $m = 2$ channels (also known as second-order PNS), a sampling interval $T_s = 1/B$, $\phi_1 = 0$, and $\phi_2 = \phi$. Due to the undersampling in each channel, aliases of the band contents tile the spectrum, so that the positive and negative images fold on each other, as visualized in Figure 6. In the



[FIG6] **Second-order PNS. The bandpass signal $x(t)$ is sampled by two rate-B uniform sequences with relative time delay ϕ. The interpolation filters cancel out the contribution of the undesired alias.**



[FIG7] **Block diagram of a time-interleaved ADC.**

frequency domain, the sample sequences (5) satisfy a linear system [23]

$$T_s\, Y_1(f) = X(f) + X(f - \beta(f)B), \tag{6a}$$

$$T_s\, Y_2(f) = X(f) + X(f - \beta(f)B)e^{-j2\pi\beta(f)\phi B} \tag{6b}$$

for $f \in \mathcal{I}$. The function $\beta(f) = -\beta(-f)$ is piecewise constant over $f \in \mathcal{I}$, indexing the aliased images. The exact levels and transitions of $\beta(f)$ depend explicitly on the band position as shown in Figure 6.

The aliases have unity weights in $y_1[n]$, whereas the time delay ϕ in $y_2[n]$ results in unequal weighting. System (6) is linearly independent as long as ϕ obeys

$$e^{-j2\pi\beta(f)\phi B} \neq 1. \tag{7}$$

Since $\beta(f)$ can take on only four distinct values within $f \in (f_l, f_u)$, there are many possible selections for ϕ that satisfy (7). Recovery of $x(t)$ is carried out by interpolation [22], [23]

$$x(t) = \sum_{n\in\mathbb{Z}} y_1[n]g_1(t - nT_s) + y_2[n]\, g_2(t - nT_s) \tag{8}$$

with bandpass filters $g_1(t), g_2(t)$, which reverse the weights in (6). These filters have frequency responses

$$G_1(f) = \frac{1}{1 - e^{-j2\pi\beta(f)\phi B}}, \quad G_2(f) = -G_1(f), \quad f \in \mathcal{I}, \tag{9}$$

as are drawn in Figure 6. In practice, these filters can be realized digitally, so that the output of Figure 6 is the Nyquist-rate sequence $x(nT)$, with $T = 1/2f_u$ equal to the Nyquist interval. Subsequently, a DAC device may interpolate the continuous signal $x(t)$.

The extension to multiband signals with N bands of individual widths B is accomplished following the same procedure using an Nth order PNS system, with delays ϕ_l, $1 \le l \le N$ [23]. Reconstruction consists of N filters, which are piecewise constant over the frequency support of $x(t)$. The indexing function $\beta(f)$ is extended to an $N \times N$ matrix $\mathbf{A}(f)$, with entries depending on ϕ_l and band locations. In general, an Nth-order PNS can resolve up to N aliases, since it provides a set of N equations. The equations are linearly independent, or solvable, if $\mathbf{A}^{-1}(f)$ exists over the entire multiband support [23]. Lin and Vaidyanathan show that the choice $\phi_l = l\phi$ renders $\mathbf{A}(f)$ a Vandermonde matrix, in which case the choice of the single delay ϕ is tractable. Bands of different widths are treated by viewing the bands as consisting of narrower intervals that are integer multiples of a common length. For example, if $N = 4$ (two transmissions) and $B_1 = k_1 B, B_2 = k_2 B$, then the equivalent model has $4(k_1 + k_2)$ bands of equal width B. This conceptual step allows to achieve the Landau rate. For technical completeness, the same solution applies to mixed rational-irrational bandwidths for an infinitesimal rate increase.

PNS VERSUS DEMODULATION

An apparent advantage of PNS over RF demodulation is that it can approach Landau's rate with no hardware components preceding the ADC device. This theoretical advantage, however, was not widely embraced by industry for acquisition of multiband inputs. In an attempt to reason this situation, we leverage practical insights from time-interleaving ADCs, a popular design topology used in high-speed converters [24]–[26].

Time-interleaved ADC technology splits the task of converting a wideband signal into M parallel branches, essentially utilizing Papoulis' theorem with a bank of time-delay elements. Each branch in the block diagram of Figure 7 introduces a time delay of ϕ_l seconds and subsequently samples $x(t - \phi_l)$ at rate $1/MT$, where $T = 1/f_{\text{NYQ}}$ is the Nyquist interval. Ideally, when $\phi_l = lT$, interleaving the M digital streams provides a sequence that coincides with the Nyquist rate samples $x(nT)$. A time-interleaving ADC consists of M separate T/H circuitries and quantizers, thereby relaxing design constraints by allowing each branch to perform the conversion task in a duration

NYQUIST AND UNDERSAMPLING ADC DEVICES

An ADC device, in the most basic form, repeatedly alternates between two states: track-and-hold (T/H) and quantization. During T/H, the ADC tracks the signal variation. When an accurate track is accomplished, the ADC holds the value steady so that the quantizer can convert the value into a finite representation. Both operations must end before the next signal value is acquired.

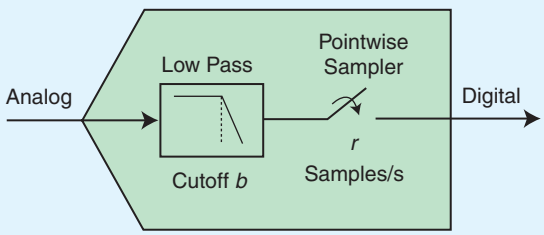
In the signal processing community, an ADC is often modeled as an ideal pointwise sampler that captures values of $x(t)$ at a constant rate of r samples per second. As with any analog circuitry, the T/H function is limited in the range of frequencies it can accept: a lowpass filter with cutoff b can be used to model the T/H capability, as depicted in Figure S1(a) [20].

In most off-the-shelf ADCs, the analog bandwidth parameter b is specified higher than the maximal sampling rate r of the device. Figure S1(b) lists example devices. When using an ADC at the Nyquist rate of the input, the filter can be omitted from the model, since the signal is bandlimited to $f_{\max} = r/2 \leq b$. In contrast, for sub-Nyquist purposes, the analog bandwidth b becomes an important factor in accurate modeling and actual selection of the ADC, since it defines the maximal input frequency that can be undersampled

$$f_{\max} \leq b. \tag{S1}$$

Typically, b specifies the -3 dB point of the T/H frequency response. Thus, if flat response in the passband is of interest, $f_{\max}$ cannot approach too close to b. For example, if $x(t)$ is a bandpass signal in the range $[600, 625]$ MHz, then undersampling at rate $f_s = 50$ MHz satisfies condition (4). In this example, while both AD9433 and AD10200 are capable of sampling at a rate $r \geq 50$ MHz, only the former is applicable due to (S1).

Undersampling ADCs have a wider spacing between consecutive samples. This advantage is translated into simplifying design constraints, especially in the duration allowed for quantization. However, regardless of the sampling rate r, the T/H stage must still hold a pointwise value of a fast-varying signal. In terms of analog bandwidth there is no substantial difference between Nyquist and undersampling ADCs; both have to accommodate the Nyquist rate of the input.



(a)

Device	Max. Rate (MSamples/s)	Analog BW (MHz)
ADS5474	400	1,440
AD12401	400	480
AD1020	105	250
AD9433	105	750

(b)

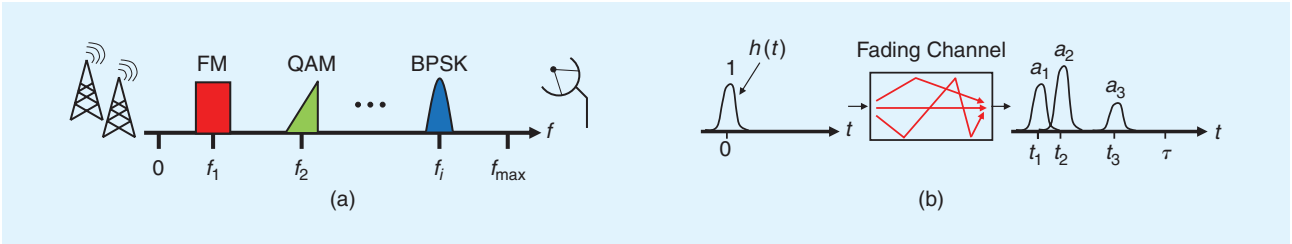
[FIG S1] **(a) A practical model for an ADC device. (b) Example devices.**

of MT seconds rather than T. While the larger duration simplifies quantization, the T/H complexity remains almost the same—it still needs to track a Nyquist-rate varying input and hold its value at a certain time point, regardless of the higher duration allocated for conversion, as explained in "Nyquist and Undersampling ADC Devices."

PNS is a degenerated time-interleaved ADC with only $m < M$ branches. This means that a PNS-based sub-Nyquist solution requires Nyquist-rate T/H circuitries, one per sampling branch. In addition to high analog bandwidth, PNS also requires compensating for imperfect production of the time delay elements. Consequently, realizing PNS in practice may not be much easier than designing an M-channel time-interleaved ADC with Nyquist-rate sampling capabilities. Thus, while time-interleaving is a popular design method for Nyquist ADCs, it may be less useful for the purpose of sub-Nyquist sampling of wideband signals with large f_{NYQ}.

More broadly, any pointwise strategy, which is applied directly on a wideband signal, has a technological barrier around the maximal rate of commercial T/H circuitry. This barrier creates an (undesired) coupling between advances in RF and ADC technologies; as transmission frequencies grow higher, a comparable speed-up of T/H bandwidth is required. With accelerated development of RF devices, a considerable gap has already been opened, rendering ADCs a bottleneck in many modern signal processing applications. In contrast, in demodulation, even though the signal is wideband, an ADC with low analog bandwidth is sufficient due to the preceding lowpass filter. RF preprocessing (mixers and filters) buffer between $x(t)$ and actual ADCs, thereby offering a scalable sampling solution, which effectively decouples T/H capabilities from dependency on the input's maximal frequency. More importantly, demodulation ensures that only in-band noise enters the system, whereas in PNS, out-of-band noise from the entire Nyquist bandwidth is aggregated.

We now turn to review sub-Nyquist techniques when the carrier frequencies are unknown, as well as low rate sampling strategies for other interesting analog models. The insights we gathered so far hint that analog preprocessing is



[FIG8] **Parts (a) and (b) show example applications of UoS modeling.**

an advantageous route towards developing efficient sub-Nyquist strategies.

UNION OF SUBSPACES

MOTIVATION

Demodulation, a classic sub-Nyquist strategy, assumes an input signal that lies in certain intervals within the Nyquist range. But, what if the input signal is not limited to a predefined frequency support, or even worse if it spans the entire Nyquist range—can we still reduce the sampling rate below Nyquist? Perhaps surprising, we shall see in the sequel that the answer is affirmative, provided that the input has additional structure we can exploit. Figure 8 illustrates two such scenarios.

Consider for example the scenario of a multiband input $x(t)$ with unknown spectral support, consisting of N frequency bands of individual widths no greater than B Hz. In contrast to the classic setup, the carrier frequencies f_i are unknown, and we are interested in sampling such multiband inputs with transmissions located anywhere below $f_{\max}$. At first sight, it may seem that sampling at the Nyquist rate $f_{\text{NYQ}} = 2f_{\max}$ is necessary, since every frequency interval below $f_{\max}$ appears in the support of some multiband $x(t)$. On the other hand, since each specific $x(t)$ in this model has structure—it fills only a portion of the Nyquist range (only NB Hz)—we intuitively expect to be able to reduce the sampling rate below f_{NYQ}.

Another interesting problem is sampling of signals that consists of several echoes of a known pulse shape, where the delays and attenuations are a priori unknown. Mathematically,

$$x(t) = \sum_{\ell=1}^{L} a_\ell\, h(t - t_\ell),\ \ t \in [0, \tau] \tag{10}$$

for some given pulse shape $h(t)$ and unknown t_ℓ, a_ℓ. Signals of this type belong to the broader family of FRI signals, originally introduced by Vetterli et al. in [27] and [28]. Echoes are encountered, for example, in multipath fading communication channels. The transmitter can assist the receiver in channel identification by sending a short probing pulse $h(t)$, based on which the receiver can resolve the fading delays t_ℓ and use this information to decode subsequent information messages. In radar applications, inputs of the form (10) are prevalent, where the delays t_ℓ correspond to the unknown locations of targets in space, while the amplitudes a_ℓ encode Doppler shifts indicating target speeds. Medical imaging techniques, e.g., ultrasound, record signals that are structured according to (10) when probing density changes in human tissue. Underwater acoustics also conform with (10). The common denominator of these applications is that $h(t)$ is a short pulse in time, so that the bandwidth of $h(t)$, and consequently that of $x(t)$, spans a large Nyquist range. Nonetheless, given the structure (10), we can intuitively expect to determine $x(t)$ from samples at the rate of innovation, namely $2L$ samples per τ, which counts the actual number of unknowns, t_ℓ, a_ℓ, $1 \le \ell \le L$ in every interval.

These examples hint at a more general notion of sub-Nyquist sampling, in which the underlying signal structure is utilized to reduce acquisition rate below the apparent input bandwidth. As a special case, this notion includes the classic settings of structure given by a predefined frequency support. To capture more general structures, we present next the union of subspace (UoS) model, originally proposed by Lu and Do in [29].

MATHEMATICAL FRAMEWORK

Denote by $x(t)$ an analog signal in the Hilbert space $\mathcal{H} = L_2(\mathbb{R})$, which lies in a parameterized family of subspaces

$$x(t) \in \mathcal{U} \triangleq \bigcup_{\lambda \in \Lambda} \mathcal{A}_\lambda, \tag{11}$$

where Λ is an index set, and each individual $\mathcal{A}_\lambda$ is a subspace of $\mathcal{H}$. The key property of the UoS model (11) is that the input $x(t)$ resides within $\mathcal{A}_{\lambda^*}$ for some $\lambda^* \in \Lambda$, but a priori, the exact subspace index λ^* is unknown. We define the dimension (or bandwidth) of $\mathcal{U}$ as the dimension of its affine hull Σ, namely the space of all linear combinations of $x(t) \in \mathcal{U}$. Typically, the union $\mathcal{U}$ has dimension that is relatively high compared with those of the individual subspaces $\mathcal{A}_\lambda$.

Multiband signals with unknown carriers f_i can be described by (11), where each $\mathcal{A}_\lambda$ corresponds to signals with specific carrier positions and the union is taken over all possible $f_i \in [0, f_{\max}]$. In this case, each $\mathcal{A}_\lambda$ has effective bandwidth NB, whereas the union $\mathcal{U}$ has $f_{\max}$ bandwidth, as follows from the definition of Σ. Similarly, echoes with unknown time delays of the form (10) correspond to L -dimensional subspaces $\mathcal{A}_\lambda$ that capture the amplitudes a_ℓ. A union over all possible delays $t_l \in [0, \tau]$ provides an efficient way to group these infinitely many subspaces to a single set

GENERALIZED SAMPLING IN UNION OF SUBSPACES

Generalized sampling theory extends upon pointwise acquisition by viewing the measurements as inner products [3]-[6], [35],

$$y[n] = \langle x(t), s_n(t) \rangle, \quad n \in \mathbb{Z}, \tag{S2}$$

between an input signal $x(t)$ and a set of sampling functions $s_n(t)$. Geometrically, the sample sequence $y[n]$ is obtained by projecting $x(t)$ onto the space

$$\mathcal{S} = \text{span}\{s_n(t) | \, n \in \mathbb{Z}\}. \tag{S3}$$

A special case is of a shift-invariant space $\mathcal{S}$ spanned by $s_n(t) = s(t-nT)$ for some generator function $s(t)$ [5]. In this scenario, (S2) amounts to filtering $x(t)$ by $s(-t)$ and taking pointwise samples of the output every T seconds. Traditional pointwise acquisition $y[n] = x(nT)$ corresponds to a shift-invariant $\mathcal{S}$ with a Dirac generator $s(t) = \delta(t)$. Multichannel sampling schemes correspond to a shift-invariant space $\mathcal{S}$ spanned by shifts of multiple generators [36], [37].

Theory and applications of subspace sampling were widely studied over the years. If $x(t)$ resides within a subspace $\mathcal{A} \subseteq \mathcal{H}$ of an ambient Hilbert space $\mathcal{H}$, then the samples (S2) determine the input whenever the orthogonal complement $\mathcal{A}^\perp$ satisfies a direct sum condition [6]

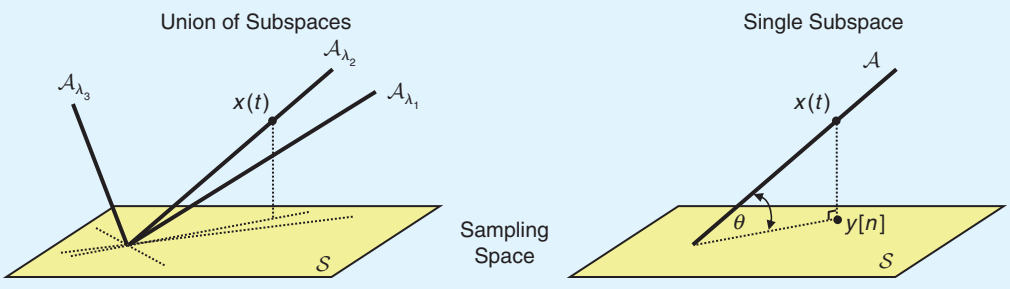
$$\mathcal{A}^\perp \oplus \mathcal{S} = \mathcal{H}. \tag{S4}$$

Reconstruction is obtained by an oblique projection [6]. Roughly speaking, in noiseless settings, perfect recovery is possible whenever the angle θ between the subspaces $\mathcal{A}, \mathcal{S}$ is different than 90°, and robustness to noise increases as θ tends to zero, as visualized in Figure S2.

When $x(t)$ lies in a union of subspaces (11), both theory and practice become more intricate. For instance, even if the angles between $\mathcal{S}$ and each of the subspaces $\mathcal{A}_\lambda$ are sufficiently small, the samples may not determine the input if several subspaces are too close to each other; see the illustration. Recent studies [29] have shown that (S2) is stably invertible if (and only if) there exist constants $0 < \alpha < \beta < \infty$ such that

$$\alpha \|x_1(t) - x_2(t)\|_{\mathcal{H}}^2 \le \|y_1[n] - y_2[n]\|_{l_2}^2 \le \beta \|x_1(t) - x_2(t)\|_{\mathcal{H}}^2, \tag{S5}$$

for every signals $x_1(t), x_2(t) \in \mathcal{A}_{\lambda_1} + \mathcal{A}_{\lambda_2}$ and for all possible pairs of λ_1, λ_2. In practice, sampling methods for specific union applications use certain hardware constraints to hint at preferred selections of stable sampling functions $s_n(t)$; see, for example, [20], [27], and [38]-[41] and other UoS methods surveyed in this review.



[FIG S2] **Geometric illustration of sampling in a single subspace versus in a union of subspaces.**

$\mathcal{U}$. The large bandwidth of $h(t)$ results in $\mathcal{U}$ with a high Nyquist bandwidth.

Union modeling sheds new light on sampling below the Nyquist rate. Sub-Nyquist in the union setting, conceptually, consists of two layers of rate reduction: from the dimensions of $\mathcal{U}$ to that of the individual subspaces $\mathcal{A}_\lambda$, and then, further reduction within the scope of a single subspace until reaching its effective bandwidth (rather than twice its highest frequency component). The second layer is essentially what is treated in the classic works surveyed earlier, which considered a single subspace defined according to a given spectral support. Eventually, the challenging part is how to design sampling strategies that combine these reduction steps and achieve the minimal rate by one conversion stage. This challenge is further discussed in "Generalized Sampling in Union of Subspaces."

The model (11) can be categorized to four types, according to the cardinality of Λ (finite or infinite) and the dimensions of the individual subspaces $\mathcal{A}_\lambda$ (finite or infinite). In the next sections, we review sub-Nyquist sampling methods for several prototype union models (categorized hereafter by the dimensions pair $\lambda - \mathcal{A}_\lambda$, where "F" abbreviates finite):

- multiband with unknown carrier positions (type F$-\infty$)
- variants of FRI models (cover two union types: ∞−F and $\infty-\infty$)
- a sparse sum of harmonic sinusoids (type F−F).

A solution for sampling and reconstruction was developed in [30] for more general F−F union structures. A special case of the F−F case is the sparsity model underlying compressed sensing [31], [32]. In this review, however, our prime focus is analog signals that exhibit infiniteness in either Λ or $\mathcal{A}_\Lambda$. A more

UNIVERSAL BOUNDS ON SUB-NYQUIST SAMPLING RATES

Sampling strategies are often compared on the basis of the required sampling rate. It is therefore instructive to compare existing strategies with the lowest sampling rate possible. For instance, the Shannon-Nyquist theorem states (and achieves) the minimal rate $2f_{\max}$ for bandlimited signals. The following results derive the lowest sub-Nyquist sampling rates for spectrally sparse signals, under either subspace or union of subspace priors.

Consider the case of a subspace model for signals that are supported on a fixed set $\mathcal{I}$ of frequencies

$$\mathcal{B}_{\mathcal{I}} = \{x(t) \in L^2(\mathbb{R}) \mid \text{supp } X(f) \subseteq \mathcal{I}\}. \tag{S6}$$

A grid $R = \{t_n\}$ of time points is called a sampling set for $\mathcal{B}_{\mathcal{I}}$ if the sequence of samples $x_R[n] = x(t_n)$ is stable, namely there exist constants $\alpha > 0$ and $\beta < \infty$ such that

$$\alpha\|x(t)-y(t)\|_{L_2}^2 \le \|x_R[n]-y_R[n]\|_{l_2}^2 \le \beta\|x(t)-y(t)\|_{L_2}^2, \quad \forall x(t), y(t) \in \mathcal{B}_{\mathcal{I}}. \tag{S7}$$

Landau [19] proved that if R is a sampling set for $\mathcal{B}_{\mathcal{I}}$ then it must have a density

$$D^-(R) \triangleq \liminf_{r\to\infty} \inf_{y\in\mathbb{R}} \frac{|R \cap [y, y+r]|}{r} \ge \text{meas}(\mathcal{I}), \tag{S8}$$

where $D^-(R)$ is the lower Beurling density and $\text{meas}(\mathcal{I})$ is the Lebesgue measure of $\mathcal{I}$. The numerator in (S8) counts the number of points from R in every interval of width r of the real axis. The Beurling density (S8) reduces to the usual concept of the average sampling rate for uniform and periodic nonuniform sampling. Consequently, for multiband signals with N bands of widths B, the minimal sampling rate is the sum of the bandwidths NB, given a fixed subspace description of known band locations.

A UoS model can describe a more general scenario, in which, a priori, only the fraction $0 < \Omega < 1$ of the Nyquist bandwidth actually occupied is assumed known but not the band locations

$$\mathcal{N}_\Omega = \{x(t) \in L^2(\mathbb{R}) \mid \text{meas}(\text{supp } X(f)) \le \Omega f_{\text{NYQ}}\}. \tag{S9}$$

A blind sampling set R for $\mathcal{N}_\Omega$ is stable if there exists $\alpha > 0$ and $\beta < \infty$ such that (S7) holds with respect to all signals from $\mathcal{N}_\Omega$. A theorem of [42] derived the minimal rate requirement for the set $\mathcal{N}_\Omega$

$$D^-(R) \ge \min\{2\Omega f_{\text{NYQ}}, f_{\text{NYQ}}\}. \tag{S10}$$

This requirement doubles the minimal sampling rate to $2NB$ for multiband signals whose band locations are unknown. It also implies that if the occupation $\Omega > 50\%$, then no rate reduction is possible.

Both minimal rate theorems are universal for pointwise sampling strategies in the sense that for any choice of a grid $R = \{t_n\}$, if the average rate is too low, particularly below (S8) or (S10), then there exist signals whose samples on R are indistinguishable. Note that both results are nonconstructive; they do not hint at a sampling strategy that achieves the minimal rate.

detailed treatment of the general union setting can be found in [33] and [34].

MULTIBAND SIGNALS WITH UNKNOWN CARRIER FREQUENCIES

UNION MODELING

A description of a multiband union can be obtained by letting $\lambda = \{f_i\}$, so that each choice of carrier positions f_i determines a single subspace $\mathcal{A}_\lambda$ in $\mathcal{U}$. In principle, f_i lies in the continuum $f_i \in [0, f_{\max}]$, resulting in union type $\infty-\infty$ containing infinitely many subspaces. In the setup we describe below a different viewpoint is used to treat the multiband model as a finite union of bandpass subspaces (type $\text{F}-\infty$), termed spectrum slices [20], [42].

In this viewpoint, the Nyquist range $[-f_{\max}, f_{\max}]$ is conceptually divided into $M = 2L + 1$ consecutive, nonoverlapping, slices of individual widths f_p, such that $Mf_p \ge 2f_{\max}$. Each spectrum slice represents a single bandpass subspace. By choosing $f_p \ge B$, we ensure that no more than $2N$ spectrum slices are active, namely contain signal energy. In this setting, there are $\binom{M}{2N}$ subspaces in $\mathcal{U}$. Dividing the spectrum to slices is only a conceptual step, which assumes no specific displacement with respect to the band positions. The advantage of this viewpoint is that switching to union type $\text{F}-\infty$ simplifies the digital reconstruction algorithms, while preserving a degree of infiniteness in the dimensions of each individual subspace $\mathcal{A}_\lambda$.

SEMIBLIND AND FULLY BLIND POINTWISE APPROACHES

Earlier approaches for treating multiband signals with unknown carriers were semiblind: a sampler design independent of band positions combined with a reconstruction algorithm that requires exact support knowledge. Herley et al. [43] and Bresler et al. [44], [45] studied multicoset sampling, a PNS grid that is a subset of the Nyquist grid, and proved that the grid points can be selected independently of the band positions. The reconstruction algorithms in [43] and [45] coincide with the nonblind PNS reconstruction algorithm of [23], for time delays ϕ_l chosen on the Nyquist grid. These works approach the Landau rate, namely NB samples/s. Other techniques targeted the rate NB by imposing alternative constraints on the input spectrum [44].

Recently, the math and algorithms for fully blind systems were developed in [39], [42], and [46]. In this setting, both sampling and reconstruction operate without knowledge of the band positions. A fundamental distinction between

nonblind or semiblind approaches to fully blind systems is that the minimal sampling rate increases to $2NB$, as a consequence of recovery that lacks knowledge on the exact spectral support. A more thorough discussion in [42] studies the differences between earlier approaches that were based on subspace modeling and the fully blind sampling methods [39], [42], [46] that are based on union modeling. "Universal Bounds on Sub-Nyquist Sampling Rates" reviews the theorems underlying this distinction. The fully blind framework developed in [42] and [46] provides reconstruction algorithms that can be combined with various sub-Nyquist sampling strategies: multicoset in [42], filter bank followed by uniform sampling in [39], and the modulated wideband converter (MWC) in [20]. In viewing our goal of bridging theory and practice, the Achilles heel of the combination with multicoset is pointwise acquisition, which enters the Nyquist-rate thru the backdoor of T/H bandwidth. As discussed earlier and outlined in "Nyquist and Undersampling ADC Devices," pointwise acquisition requires an ADC device with Nyquist-bandwidth T/H circuitry. The filter-bank approach is part of a general framework developed in [39] for treating analog signals lying in a sparse-sum of shift-invariant (SI) subspaces, which includes multiband with unknown carriers as a special case. The filters and ADCs, however, also require Nyquist-rate bandwidth, in this setting.

In the next section, we describe the MWC strategy, which utilizes the principles of the fully blind sampling framework, and also results in a hardware-efficient sub-Nyquist strategy that does not suffer from analog bandwidth limitations of T/H technology. In essence, the MWC extends conventional I/Q demodulation to multiband inputs with unknown carriers, and as such it also provides a scalable solution that decouples undesired RF-ADC dependencies. The combination of hardware-efficient sampler with fully blind reconstruction effectively satisfies the wish list of Table 1.

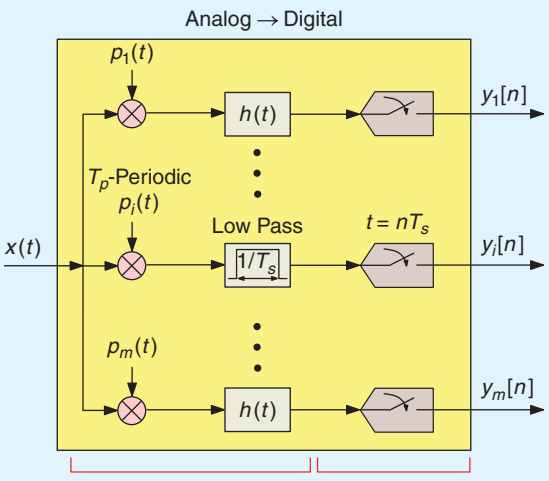
MODULATED WIDEBAND CONVERTER

The MWC [20] combines the advantages of RF demodulation and the blind recovery ideas developed in [42], and allows sampling and reconstruction without requiring knowledge of the band locations. To circumvent analog bandwidth issues in the ADCs, an RF front end mixes the input with periodic waveforms. This operation imitates the effect of delayed undersampling, specifically folding the spectrum to baseband with different weights for each frequency interval. In contrast to undersampling (or PNS), aliasing is realized by RF components rather than by taking advantage of the T/H circuitry. In this way, bandwidth requirements are shifted from ADC devices to RF mixing circuitries. The key idea is that periodic mixing serves another goal—both the sampling and reconstruction stages do not require knowledge of the carrier positions.

Before describing the MWC system that is depicted in Figure 9, we point out several properties of this approach. The system is modular; Sampling is carried out in independent channels, so that the rate can be adjusted to match the requirements of either a traditional subspace model or the more challenging union of subspace prior. It can also scale up to the Nyquist rate to support the standard Shannon-Nyquist bandlimited prior. The reconstruction algorithm that appears in Figure 10 has several functional blocks: detecting the spectral support through a computationally light optimization problem, signal recovery, and information extraction. Support detection, the heart of this digital algorithm, is carried out whenever the carrier locations vary. The rest of the digital computations are simple and performed in real time. In addition, the recovery stage outputs baseband samples of $I(t), Q(t)$. This enables a seamless interface to existing DSP algorithms with sub-Nyquist processing rates, as could have been obtained by classic demodulation had the carriers f_i been known. We now elaborate on each part of this strategy.

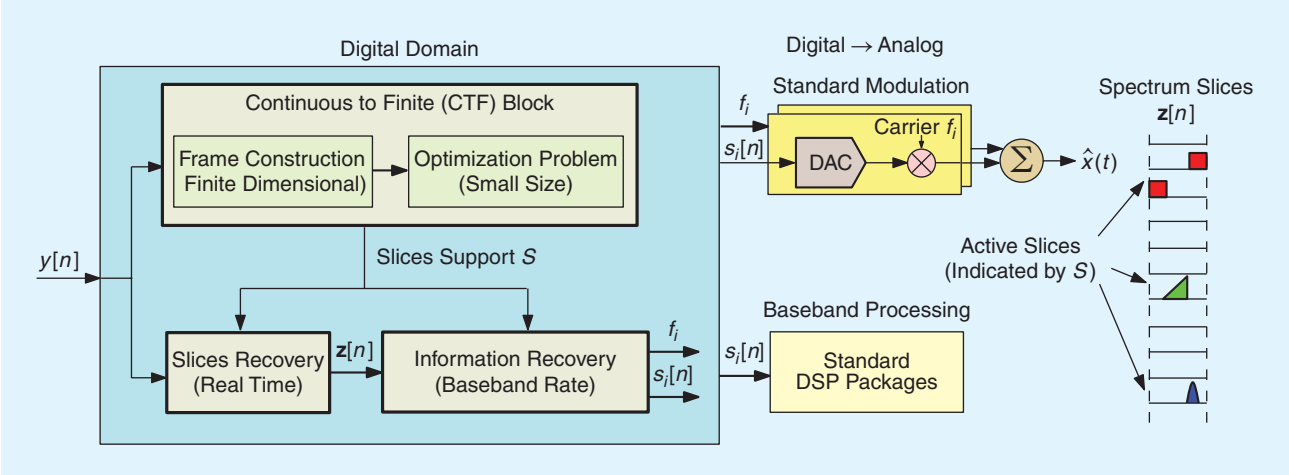
SUB-NYQUIST SAMPLING SCHEME

The conversion from analog to digital consists of a front end of m channels, as depicted in Figure 9. In the ith channel, $x(t)$ is multiplied by a periodic waveform $p_i(t)$ with period $T_p = 1/f_p$, lowpass filtered by $h(t)$, and then sampled at rate $f_s = 1/T_s$. The figure lists basic and advanced configurations. To simplify, we concentrate on the theory underlying the basic version, in which the sampling interval T_s equals the aliasing period T_p, each channel samples at rate $f_s \geq B$ and the number of



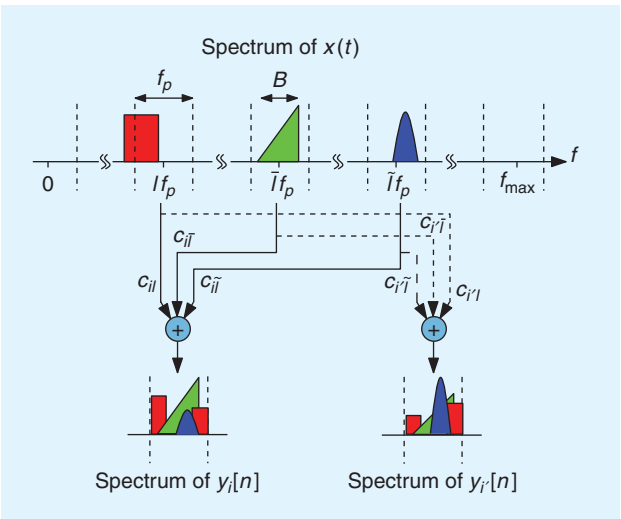
Parameter/Configuration	Basic	Advanced
# Channels m	$\geq 4N$	$\geq 2N/q$
# Periodic Generators	$=m$	1
Aliasing Rate $1/T_p$	$\geq B$	$\geq B$
Cutoff/Channel Rate $1/T_s$	$=1/T_p$	$=q/T_p$
Total Rate	Approach $4NB$	Approach $2NB$

[FIG9] (a) Block diagram of the modulated wideband converter [20] (figure courtesy of the IEEE). Part (b) lists the parameters choice of the basic and advanced MWC configurations. Adapted from [47] (figure courtesy of IET).



[FIG10] **Block diagram of recovery and processing stages of the modulated wideband converter.**

hardware branches $m \geq 2N$, so that the total sampling rate can be as low as $2NB$. These choices stem from necessary and sufficient conditions derived in [20] on the required sampling rate mf_s to allow perfect reconstruction. If the input's spectral support is known, then the same conditions translate to a similar parameter choice with half the number of channels, resulting in a total sampling rate as low as NB. Thus, although the MWC does not take pointwise values of $x(t)$, its optimal sampling rate coincides with the lowest possible rates by pointwise strategies, which are discussed in "Universal Bounds on Sub-Nyquist Sampling Rates." Advanced configurations enable additional hardware savings by collapsing the number of branches m by a factor of q at the expense of increasing the sampling rate of each channel by the same factor, ultimately enabling a single-channel sampling system [20]. This technique is also briefly reviewed in the next subsection.



[FIG11] **The spectrum slices from $x(t)$ are overlayed in the spectrum of the output sequences $y_i[n]$. In the example, channels i and i' realize different linear combinations of the spectrum slices centered around lf_p, $\tilde{l}f_p$, $\tilde{\tilde{l}}f_p$. For simplicity, the aliasing of the negative frequencies is not drawn. Adapted from [47] (figure courtesy of IET).**

The choice of periodic waveforms $p_i(t)$ becomes clear once analyzing the effect of periodic mixing. Each $p_i(t)$ is periodic, and thus has a Fourier expansion

$$p_i(t) = \sum_{l=-\infty}^{\infty} c_{il} e^{j2\pi f_p l t}. \tag{12}$$

Denote by $z_l[n]$ the sequence that would have been obtained by mixing $x(t)$ with $e^{j2\pi f_p l t}$, filtering by $h(t)$ and sampling every T seconds. By superposition, mixing $x(t)$ by the sum in (12) outputs $y_i[n]$, which is a linear combination of the $z_l[n]$ sequences according to the Fourier coefficients c_{il} of $p_i(t)$. Figure 11 visualizes the effect of mixing with periodic waveforms, where each sequence $z_l[n]$ corresponds to a spectrum slice of $x(t)$ positioned around lf_p. Mathematically, the analog mixture boils down to the linear system [20]

$$\mathbf{y}[n] = \mathbf{C}\mathbf{z}[n], \tag{13}$$

where the vector $\mathbf{y}[n] = [y_1[n], \ldots, y_m[n]]^T$ collects the measurements at $t = nT_s$. The matrix $\mathbf{C}$ contains the coefficients c_{il} and $\mathbf{z}[n]$ rearranges the values of $z_l[n]$ in vector form.

To enable aliasing of spectrum slices up to the maximal frequency $f_{\max}$, the periodic functions $p_i(t)$ need to have Fourier coefficients c_{il} with nonnegligible amplitudes for all $-L \leq l \leq L$, such that $Lf_p \geq f_{\max}$. In principle, every periodic function with high-speed transitions within the period T_p can be appropriate. One possible choice for $p_i(t)$ is a sign-alternating function, with $M = 2L + 1$ sign intervals within the period T_p [20]. Popular binary patterns, e.g., the Gold or Kasami sequences, are especially suitable for the MWC [38].

HARDWARE-EFFICIENT REALIZATION

A board-level hardware prototype of the MWC is reported in [47]. The hardware specifications cover 2 GHz Nyquist-rate inputs with spectral occupation up to $NB = 120$ MHz. The

sub-Nyquist rate is 280 MHz. Photos of the hardware appear in Figure 12.

To reduce the number of analog components, the hardware realization incorporates an advanced MWC configuration [20]. In this version

- a collapsing factor $q = 3$ results in $m = 4$ hardware branches with individual sampling rates $1/T_s = 70$ MHz
- a single shift-register generates periodic waveforms for all hardware branches.

Further technical details on this representative hardware exceed the level of practice we are interested in here, though we emphasize below a few conclusions that connect back to the theory.

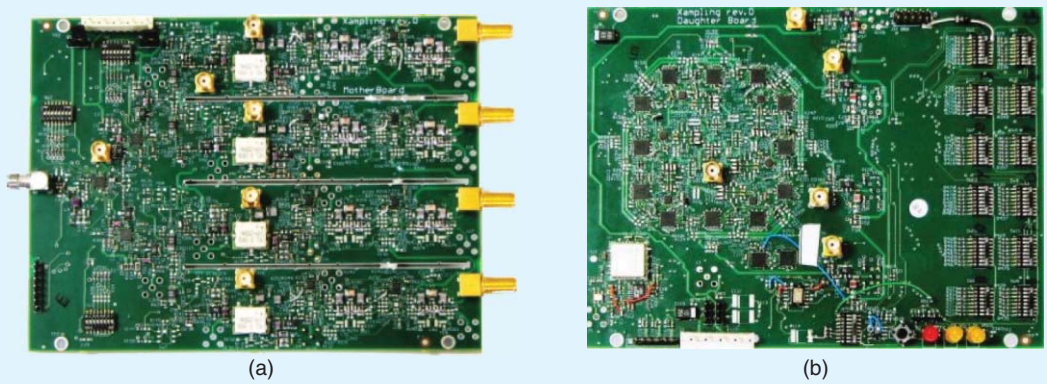
The Nyquist burden always manifests itself in some part of the design. For example, in pointwise methods, implementation requires ADC devices with Nyquist-rate front-end bandwidth. In other approaches [41], [48], which we discuss in the sequel, the computational loads scale with the Nyquist rate, so that an input with 1 MHz Nyquist rate may require solving linear systems with 1 million unknowns. Example hardware realizations of these techniques [49] treat signals with Nyquist rate up to 800 kHz. The MWC shifts the Nyquist burden to an analog RF preprocessing stage that precedes the ADC devices. The motivation behind this choice is to enable capturing the largest possible range of input signals, since, in principle, when the same technology is used by the source and sampler, this range is maximal. In particular, as wideband multiband signals are often generated by RF sources, the MWC framework can treat an input range that scales with any advance in RF technology.

While this explains the choice of RF preprocessing, the actual sub-Nyquist circuit design can be quite challenging and call for nonordinary solutions. To give a taste of circuit challenges, we briefly consider two design problems that are studied in detail in [47]. Low-cost analog mixers are typically specified for a pure sinusoid in their oscillator port, whereas the periodic mixing requires simultaneous mixing with the many sinusoids comprising $p_i(t)$, which creates nonlinear distortions and complicates the gain selections along the RF path. In [47], special power circuitries that are tailored for periodic mixing were inserted before and after the mixer. Another circuit challenge pertains to generating $p_i(t)$ with 2 GHz alternation rates. The strict timing constraints involved in this logic are eliminated in [47] by operating commercial devices beyond their datasheet specifications.

IN PRINCIPLE, WHEN THE SAME TECHNOLOGY IS USED BY THE SIGNAL SOURCE AND SAMPLER, THE RANGE OF POSSIBLE INPUT SIGNALS IS MAXIMIZED.

Going back to a high-level practical viewpoint, besides matching the source and sampler technology and addressing circuit challenges, an important point is to verify that the recovery algorithms do not limit the input range through constraints they impose on the hardware. In the MWC case, periodicity of the waveforms $p_i(t)$ is important since it creates the aliasing effect with the Fourier coefficients c_{il} in (12). The hardware implementation and experiments in [47] demonstrate that the appearance in time of $p_i(t)$ is irrelevant as long as periodicity is maintained. A video recording of hardware experiments and additional documentation for the MWC hardware is available in [85]. This property is crucial, since precise sign alternations at high speeds of 2 GHz are difficult to maintain, whereas simple hardware wirings ensure periodicity, specifically that $p_i(t) = p_i(t + T_p)$ for every $t \in \mathbb{R}$. The recent work [50] provides digital compensation for nonflat frequency response of $h(t)$, assuming slight oversampling to accommodate possible design imperfections, similarly to oversampling solutions in Shannon-Nyquist sampling.

Noise is inevitable in practical measurement devices. A common property of many existing sub-Nyquist methods, including PNS sampling, MWC, and the methods of [41] and [48] is that they aggregate wideband noise from the entire Nyquist range, as a consequence of treating all possible



[FIG12] A hardware realization of the MWC consisting of two circuit boards. Part (a) implements $m = 4$ sampling channels, whereas (b) provides four sign-alternating periodic waveforms of length $M = 108$, derived from a single shift-register. Adapted from [47] (figure courtesy of IET).

SPARSE SOLUTIONS OF UNDERDETERMINED LINEAR SYSTEMS

A famous riddle is as follows: "You are given a balanced scale and 12 coins, one of which is counterfeit. The counterfeit weighs less or more than the other coins. Determine the counterfeit in three weighings, and whether it is heavier or lighter." This riddle captures the essence of sparsity priors. While there are multiple unknowns (the weights of the 12 coins), far fewer measurements (only three) are required to determine low-dimensional information of interest (the relative weight of the counterfeit coin). Several "12 coins" solutions (widely available online) are based on three rounds of comparing weights of two groups of four coins each, followed by a sort of combinatorial logic that indicates the counterfeit coin.

Sparse solutions of underdetermined linear systems extend the principle underlying the above riddle. The influential works by Donoho [31] and Candès et al. [32] paved the way to CS, an emerging field in which problems of this type are widely studied. Mathematically, consider the linear system

$$\mathbf{y} = \mathbf{C}\mathbf{z}, \tag{S11}$$

with the $m \times M$ matrix $\mathbf{C}$ having fewer rows than columns, i.e., $m < M$. Since $\mathbf{C}$ has a nontrivial null space, there are infinitely many candidates $\mathbf{z}$ satisfying (S11). The goal of CS is to find a sparse $\mathbf{z}$ among these solutions, in other words, a vector $\mathbf{z}$ that contains only few nonzero entries. A basic result in the field [53] shows that (S11) has a unique sparse solution if

$$\|\mathbf{z}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu}\right), \quad \mu \triangleq \max_{i \neq j} \frac{\langle \mathbf{C}_i, \mathbf{C}_j \rangle}{\|\mathbf{C}_i\| \, \|\mathbf{C}_j\|}, \tag{S12}$$

where $\|\mathbf{z}\|_0$ counts the number of nonzeros in $\mathbf{z}$, and $\|\mathbf{C}_i\|$ denotes the Euclidian norm of the ith column $\mathbf{C}_i$. The unique sparse solution can be found via the minimization program,

$$\min_{\mathbf{z}} \|\mathbf{z}\|_0 \ \text{ s.t. } \ \mathbf{y} = \mathbf{C}\mathbf{z}. \tag{S13}$$

Similar to the riddle, program (S13) is NP-hard with combinatorial complexity.

The CS literature provides polynomial-time techniques for sparse recovery, which coincide with the sparse $\mathbf{z}$, under various conditions on the matrix $\mathbf{C}$. A popular alternative to (S13) is solving the convex program

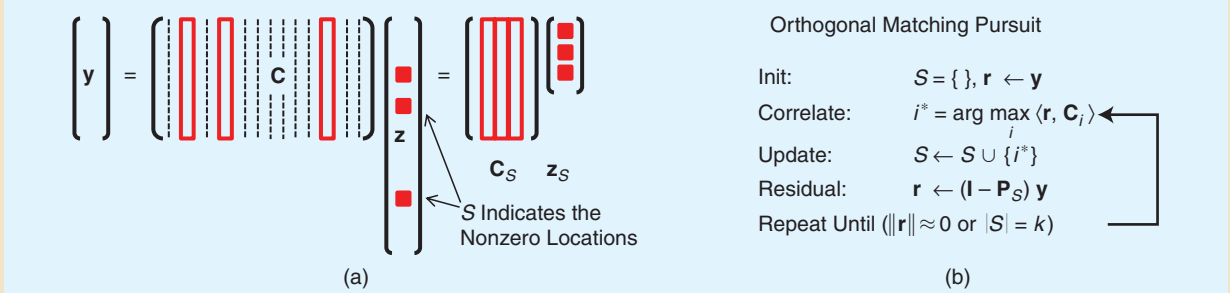
$$\min_{\mathbf{z}} \|\mathbf{z}\|_1 \ \text{ s.t. } \ \mathbf{y} = \mathbf{C}\mathbf{z}, \tag{S14}$$

where the norm $\|\mathbf{z}\|_1$ sums the magnitudes of the entries. Convex variants of (S14) include penalizing terms that account for additive noise. Another approach to sparse recovery are greedy algorithms, which iteratively recover the nonzero locations. For example, orthogonal matching pursuit (OMP) [54] iteratively identifies a single support index. A residual vector $\mathbf{r}$ contains the part of $\mathbf{y}$ that is not spanned by the currently recovered index set. In OMP, an orthogonal projection $\mathbf{P}_S\mathbf{y}$ is computed in every iteration, as described in Figure S3. Various greedy approaches are modifications of the main OMP steps.

The procedure repeats until the location set S reaches a predefined cardinality or when the residual r is sufficiently small. Upon termination, the nonzero values $\mathbf{z}_S$ are computed by pseudoinversion of the relevant column subset $\mathbf{C}_S$

$$\mathbf{z}_S = \mathbf{C}_S^{\dagger}\mathbf{y} = (\mathbf{C}_S^T\mathbf{C}_S)^{-1}\mathbf{C}_S^T\mathbf{y}. \tag{S15}$$

Convex and greedy methods have also been proposed to account for joint sparsity, in which case the unknown is a matrix $\mathbf{Z}$ having only a few nonidentically zero rows [30], [46], [55]–[59]. The special issue of *IEEE Signal Processing Magazine* from March 2008 and book [60] provide a comprehensive review of this field [61].



[FIG S3] **(a) An underdetermined system with a sparse solution vector. (b) Example sparse recovery algorithm.**

spectral supports. The digital reconstruction algorithm we outline in the next subsection partially compensates for this noise enhancement for PNS/MWC by digital denoising. Another route to noise reduction can be careful design of the sequences $p_i(t)$. However, noise aggregation remains a practical limitation of all current sub-Nyquist techniques.

RECONSTRUCTION ALGORITHM

The digital reconstruction algorithm encompasses three stages that appear in Figure 10:

1) A block named continuous-to-finite (CTF) constructs a finite-dimensional frame (or basis) from the samples, from which a small-size optimization problem is formulated.

The solution of that program indicates those spectrum slices that contain signal energy. The CTF outputs an index set S of active slices. This block is executed on initialization or when the carrier frequencies change.

2) A single matrix-vector multiplication, per instance of $\mathbf{y}[n]$, recovers the sequences $z_l[n]$ containing signal energy, as indicated by $l \in S$.

3) A digital algorithm estimates f_i and (samples of) the baseband signals $I(t)$, $Q(t)$ of each information band.

In addition to DSP, analog recovery of $x(t)$ is obtained by remodulating the quadrature signals $I(t)$, $Q(t)$ on the estimated carriers f_i according to (3). Analog back end employs customary components, DACs and modulators, to recover $x(t)$.

To understand the recovery flow, we begin with the linear system (13). Due to the sub-Nyquist setup, the matrix $\mathbf{C}$ in (13) has dimension $m \times M$, such that $m < M$. In other words, $\mathbf{C}$ is rectangular and (13) has fewer equations than the dimension M of the unknown $\mathbf{z}[n]$. Fortunately, the multiband prior in accordance with the choice $f_p \geq B$ ensures that at most $2N$ sequences $z_l[n]$ contains signal energy [20]. Therefore, for every time point n, the unknown $\mathbf{z}[n]$ is sparse with no more than $2N$ nonzero values. Solving for a sparse vector solution of an underdetermined system of equations has been widely studied in the literature of compressed sensing (CS). "Sparse Solutions of Underdetermined Linear Systems" summarizes relevant CS theorems and algorithms.

Recovery of $\mathbf{z}[n]$ using any of the existing sparse recovery techniques is inefficient, since the sparsest solution $\mathbf{z}[n]$, even if obtained by a polynomial-time CS technique, is computed independently for every n. Instead, the CTF method that was developed in [42] and [46] exploits the fact that the bands occupy continuous spectral intervals. This analog continuity boils down to $\mathbf{z}[n]$ having a common nonzero location set S over time. To take advantage of this joint sparsity, the CTF builds a frame (or a basis) from the measurements using, for example,

$$\mathbf{y}[n] \xrightarrow{\text{Frame construct}} \mathbf{Q} = \sum_n \mathbf{y}[n]\,\mathbf{y}^H[n] \xrightarrow{\text{Decompose}} \mathbf{Q} = \mathbf{V}\mathbf{V}^H, \qquad (14)$$

where the (optional) decomposition allows to combat noise. The finite dimensional system

$$\mathbf{V} = \mathbf{C}\mathbf{U}, \qquad (15)$$

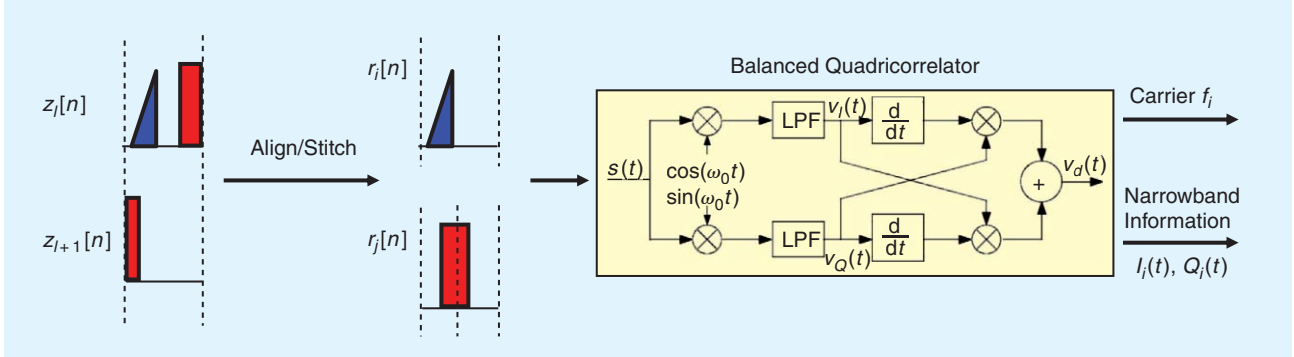
is then solved for the sparsest matrix $\mathbf{U}$ with minimal number of nonidentically zero rows; example techniques are referenced in "Sparse Solutions of Underdetermined Linear Systems." The important observation is that the indices of the nonzero rows in $\mathbf{U}$, denoted by the set S, coincide with the locations of the spectrum slices that contain signal energy [42]. This property holds for any choice of matrix $\mathbf{V}$ in (15) whose columns span the measurements space $\{\mathbf{y}[n]\}$. The CTF effectively locates the signal energy at a spectral resolution of f_p. Once S is found, $\mathbf{z}[n]$ are recovered by a matrix-vector multiplication; see (S15) in "Sparse Solutions of Underdetermined Linear Systems." Since all CTF operations are executed only once (or when carrier frequencies change), in steady state, the reconstruction runs in real time, specifically a single matrix-vector multiplication (S15) per measurement $\mathbf{y}[n]$.

SUB-NYQUIST BASEBAND PROCESSING

Software packages for DSP expect baseband inputs, specifically the information signals $I(t)$, $Q(t)$ of (3), or equivalently their uniform samples at the narrowband rate. These inputs are obtained by classic demodulation when the carrier frequencies are known. A digital algorithm developed in [51] translates the sequences $\mathbf{z}[n]$ to the desired DSP format with only lowrate computations, enabling smooth interfacing with existing DSP software packages.

The input to the algorithm are the sequences $\mathbf{z}[n]$ corresponding to the spectrum slices of $x(t)$. In general, as depicted in Figure 13, a spectrum slice may contain more than a single information band. The energy of a band of interest may also split between adjacent slices. To correct for these two effects, the algorithm performs the following actions:

1) Refine the coarse support estimate S to the actual band edges, using power spectral density estimation.



[FIG13] The flow of information extraction begins with detecting the band edges. The slices are filtered, aligned and stitched appropriately to construct distinct quadrature sequences $r_i[n]$ per information band. The balanced quadricorrelator finds the carrier f_i and extracts the narrowband information signals.

2) Separate bands occupying the same slice to distinct sequences $r_i[n]$. Stitch together energy that was split between adjacent slices.

3) Apply a common carrier recovery technique, the balanced quadricorrelator [52], on $r_i[n]$. This step estimates the carrier frequencies f_i and outputs uniform samples of the narrowband signals $I(t)$, $Q(t)$.

The baseband processing algorithm renders the MWC compliant with the high-level architecture of Figure 2 depicted in the beginning of the article. The digital computations of the MWC (CTF, spectrum slices recovery, and baseband processing) lie within the digital core that enables DSP and assist continuous reconstruction.

ADAPTIVE SOLUTIONS

We conclude this section with a brief discussion on a potential adaptive strategy for multiband sampling. An adaptive system may scan the spectrum for the frequencies f_i prior to sampling, and then employ classic solutions, e.g., demodulation or PNS, for the actual conversion to digital. This approach requires a wideband analog spectrum scanner that can be hardware excessive and time consuming; cf. [51]. During that time, signal acquisition is idle, thereby precluding reconstruction of potentially valuable data. The fact that f_i are unknown a priori and are learned while the system is running has additional implications on the hardware. For example adaptive demodulation requires a local oscillator tunable over the entire wideband range, so that it can generate a sinusoid at any identified f_i in $[0, f_{\max}]$. In PNS techniques, the sampling grid needs to be designed at run time, especially after f_i are determined, as evident from conditions (4)–(7) and Figures 5 and 6. Nonetheless, where applicable, adaptive solutions may be another venue for sub-Nyquist sampling. A prominent advantage of adaptive demodulation is that only in-band noise enters the system.

SIGNALS WITH FINITE RATE OF INNOVATION

PERIODIC TIME-DELAY MODEL

Vetterli et al. [27], [62] coined the FRI terminology for signals that are determined by a finite number L of unknowns, referred to as innovations, per time interval τ. Bandlimited signals, for example, have $L = 1$ innovations per Nyquist interval $\tau = 1/f_{\text{NYQ}}$. The most studied FRI model is that of (10), in which there are $2L$ innovations: unknown delays t_ℓ and attenuations a_ℓ of L copies of a given pulse shape $h(t)$ [27], [28], [40], [62]–[64]. As outlined earlier, the sub-Nyquist goal in this setting is to determine $x(t)$ from about $2L$ samples per τ, rather than sampling at the high rate that stems from the bandwidth of $h(t)$. In what follows, we consider a simple version of (10) with a periodic input, $x(t) = x(t + \tau)$, so that the echoes pattern, i.e., t_ℓ and a_ℓ, repeats every τ seconds. Each possible choice of delays $\{t_\ell\}$ leads to a different L-dimensional subspace of signals $\mathcal{A}_\lambda$, spanned by the functions $\{h(t-t_\ell)\}$. Since the delays lie on the continuum $t_\ell \in [0, \tau]$, the model (10) corresponds to an infinite union of finite dimensional subspaces (type $\infty-$F). We first describe the sub-Nyquist principles for this periodic version, and then outline other variants of FRI signals and sampling strategies.

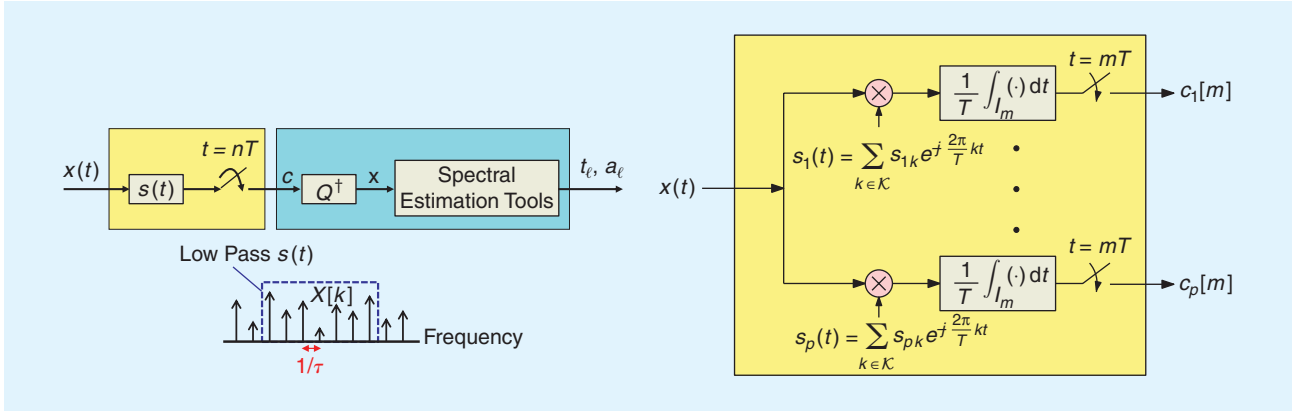
SUB-NYQUIST SAMPLING SCHEME

The key enabling sub-Nyquist sampling in the FRI setting is in identifying the connection to a standard problem in signal processing: retrieval of the frequencies and amplitudes of a sum of sinusoids. The Fourier series coefficients $X[k]$ of the periodic pulse stream $x(t)$ can be shown to equal a sum of complex exponentials, with amplitudes $\{a_\ell\}$, and frequencies directly related to the unknown time-delays [27]

$$X[k] = \frac{1}{\tau}\int_0^\tau x(t)e^{-j2\pi kt/\tau}\mathrm{d}t = \frac{1}{\tau}H(2\pi \mathrm{k}/\tau)\sum_{\ell=1}^{L} a_\ell e^{-j2\pi k t_\ell/\tau}, \quad (16)$$

where $H(\omega)$ is the Fourier transform of the pulse $h(t)$. Once the coefficients $X[k]$ are known, the delays and amplitudes can be found using standard tools developed in the context of array processing and spectral estimation [27], [65]. Therefore, the goal is to design a sampling scheme from which $X[k]$ can be determined.

Figure 14 depicts two sampling strategies to obtain $X[k]$. In the single-channel version, the input is filtered by $s(t)$ and then sampled uniformly every T seconds. If $s(t)$ is designed to capture a set $\mathbf{x}$ of $M \geq 2L$ consecutive coefficients $X[k]$ and



[FIG14] Parts (a) and (b) show single and multichannel sampling schemes for time-delay FRI models.

zero out the rest, then the vector $\mathbf{x}$ of Fourier coefficients can be obtained from the samples [63]

$$\mathbf{x} = \mathbf{S}^{-1}\,\mathrm{DFT}\{\mathbf{c}\}, \tag{17}$$

where $\mathbf{S}$ is an $M \times M$ diagonal matrix with kth entry $S^*(2\pi k/\tau)$ for all k in the filter's passband, and $\mathbf{c}$ collects M uniform samples in a time duration τ. One way to capture M coefficients $X[k]$ is by choosing a lowpass $s(t)$ with an appropriate cutoff [27]. A more general condition on the sampling kernel $s(t)$ is that its Fourier transform $S(\omega)$ satisfies [63]

$$S(\omega) = \begin{cases} 0 & \omega = 2\pi k/\tau,\ k \notin \mathcal{K} \\ \text{nonzero} & \omega = 2\pi k/\tau, k \in \mathcal{K} \\ \text{arbitary} & \text{otherwise,} \end{cases} \tag{18}$$

where $\mathcal{K}$ is a set of $M \geq 2L$ consecutive indices such that $H(2\pi k/\tau) \neq 0$ for all $k \in \mathcal{K}$. Practical (real-valued) kernels $s(t)$ have conjugate symmetric transform $S(\omega)$ and thus necessitate selecting odd M, in which case the minimal number of samples is $2L + 1$.

A special class of filters satisfying (18) are sum of sincs (SoS) in the frequency domain [63], which lead to compactly supported filters in the time domain; this property becomes crucial in other variants of FRI models we survey below. As the name hints, SoS filters are given in the Fourier domain by

$$G(\omega) = \frac{\tau}{\sqrt{2\pi}} \sum_{k\in\mathcal{K}} b_k \,\mathrm{sinc}\left(\frac{\omega}{2\pi/\tau} - k\right), \tag{19}$$

where $b_k \neq 0, k \in \mathcal{K}$. It is easy to see that this class of filters satisfies (18) by construction. Switching to the time domain

$$g(t) = \mathrm{rect}\left(\frac{t}{\tau}\right) \sum_{k\in K} b_k e^{j2\pi kt/\tau}, \tag{20}$$

which is clearly a time compact filter with support τ. For the special case in which $\mathcal{K} = \{-p, \ldots, p\}$ and $b_k = 1$,

$$g(t) = \mathrm{rect}\left(\frac{t}{\tau}\right) \sum_{k=-p}^{p} e^{j2\pi kt/\tau} = \mathrm{rect}\left(\frac{t}{\tau}\right) D_p(2\pi t/\tau), \tag{21}$$

where $D_p(t)$ denotes the Dirichlet kernel.

An alternative multichannel sampling system was proposed in [64]. The system, depicted in Figure 14(b), is conceptually constructed in two steps. First, M analog branches are used to compute $X[k]$ directly from $x(t)$ according to (16): modulation by $e^{-j2\pi kl/\tau}$ and integration over τ. For practical reasons, generating M complex sinusoids at different frequencies can be hardware excessive. Therefore, the second step replaces mixing with individual sinusoids by $x(t)s_i(t)$, with mixing waveforms $s_i(t)$ consisting of a linear combination of $|\mathcal{K}|$ complex sinusoids. The advantage is that $s_i(t)$ can be efficiently generated by proper (lowpass) filtering of periodic waveforms. The periodic waveforms themselves can be generated from a single clock source [47]. Interestingly, the MWC hardware prototype, whose boards appear in Figure 12, functions as a generic sub-Nyquist platform; it can also be used for reduced-rate sampling of FRI models [66]. In the digital domain, $X[k]$ are computed from samples of the linear mixtures $x(t)s_i(t)$.

RECONSTRUCTION ALGORITHM

Given a vector $\mathbf{x}$ of coefficients $X[k]$, solving for t_ℓ, a_ℓ from (16) is tantamount to recovering L frequencies and amplitudes in a sum of complex exponentials. A variety of methods for that problem have been proposed; see [65] for a comprehensive review. Below we outline the annihilating filter method that is used in [27], as it allows recovery from the critical rate of $2L/\tau$.

The key ingredient of the method is a digital filter $A[k]$, whose z-transform

$$A(z) = \sum_{k=0}^{L} A[k]\, z^{-k} = A[0] \prod_{l=1}^{L} (1 - e^{-j2\pi t_\ell/\tau} z^{-1}) \tag{22}$$

has zeros at the L fundamental frequencies $e^{j2\pi t_\ell/\tau}$. Convolving $A[k]$ with the coefficients $X[k]$, has an annihilating effect, namely returns zero, since each of the frequencies in $X[k]$ is canceled out by the relevant zero of $A(z)$. The idea is therefore to construct $A[k]$ and then factorize its roots to recover the fundamental frequencies, which imply t_ℓ. In turn, the amplitudes a_ℓ are found by standard linear regression tools. The annihilating filter $A[k]$ is computed from the set of constraints [27], [65]

$$\begin{bmatrix} X[0] & X[-1] & \cdots & X[-L] \\ X[1] & X[0] & \cdots & X[-(L-1)] \\ \vdots & \vdots & \ddots & \vdots \\ X[L] & X[L-1] & \cdots & X[0] \end{bmatrix} \begin{pmatrix} A[0] \\ A[1] \\ \vdots \\ A[L] \end{pmatrix} = \mathbf{0}. \tag{23}$$

Without loss of generality $A[0] = 1$ [constant scaling does not affect the roots in (22)], so that (23) determines the annihilating filter, and consequently $\{t_\ell\}_{\ell=1}^{L}$, from as low as $2L$ values of $X[k]$. As explained before, a single-channel real-valued kernel $s(t)$ requires a minimal number of samples equal to $M = 2L + 1$.

FINITE-DURATION FRI MODELS

While periodic streams are mathematically convenient, finite pulse streams of the form (10) are ubiquitous in real-world applications. For example, in ultrasound imaging, there are finitely many echoes reflected from the tissue. Radar techniques determine target locations based on echoes of a transmitted pulse, where again finitely many echoes are used. A finite-duration FRI input $x(t)$ coincides with its periodized version $\sum_{k\in\mathbb{Z}} x(t + k\tau)$ on the observation interval $[0, \tau]$. Thus, ultimately, we would like to utilize the previous sampling techniques and algorithms on that interval. The difficulty is, however, that a simple lowpass kernel $s(t)$ has infinite time support, which lasts effectively beyond the time interval $[0, \tau]$, to the point where $x(t)$ differs from its periodized

version. A more localized Gaussian sampling kernel was proposed in [27]; however, this method is numerically unstable since the samples are multiplied by a rapidly diverging or decaying exponent. Compactly supported sampling kernels based on splines were studied in [28] for certain classes of pulse shapes. These kernels enable computing moments of the signal rather than its Fourier coefficients, which are then processed in a similar fashion to obtain t_ℓ, a_ℓ.

An alternative approach is to exploit the compact support of SoS filters [63]. Since (20) is compactly supported in time by construction, the values of $x(t)$ beyond the filter support are of no interest. In particular, $x(t)$ may be zero in that range. Therefore, when using SoS filters, periodic and finite-duration FRI models are essentially treated in the same fashion. This approach exhibits superior noise robustness when compared to the Gaussian and spline methods, and can be used for stable reconstruction even for very high values of L, e.g., $L = 100$. Potential applications are ultrasound imaging [63], radar [67], and Gabor analysis in the Doppler plane [68]. Multichannel sampling, according to Figure 14, can be more efficient for implementation since accurate delay elements are avoided. The parallel scheme enjoys similar robustness to noise and allows approaching the minimal innovation rate. It is also applicable in cases of infinite pulse streams, as we discuss next.

INFINITE PULSE STREAM

The model of (10) can be further extended to an infinite stream case, in which

$$x(t) = \sum_{l\in\mathbb{Z}} a_\ell h(t-t_\ell), \quad t_\ell, a_\ell \in \mathbb{R}. \tag{24}$$

Both [28] and [63] exploit the compact support of their sampling filters and show that under certain conditions, the infinite stream may be divided into a series of finite-duration FRI problems, which are each solved independently using the previous algorithms. Since proper spacing is required between the finite streams to ensure up to L pulses within the support of the sampling kernel, the rate is increased beyond minimal. In [28], the rate scales with L^2, whereas in [63] the rate requirement is improved to about $6L$, specifically a small constant factor from the rate of innovation. A multichannel approach for the infinite model was first considered for Dirac streams, where a successive chain of integrators allows obtaining moments of the signal [69]. Exponential filters were used in [70] for the same model of Dirac streams. Unfortunately, both methods are sensitive in the presence of noise and for large values of L [64]. A simple sampling and reconstruction scheme consisting of two R-C circuit channels was presented in [71] for the special case where there is no more than one Dirac per sampling period. The system of Figure 14 can treat a broader class of infinite pulse streams, while being much more stable [64]. It exhibits superior noise robustness over the integrator chain method [69] and allows for more general compactly supported pulse shapes.

SEQUENCES OF INNOVATIONS

A special case of (24) is when the time delays repeat periodically (but not the amplitudes), resulting in

$$x(t) = \sum_{n\in\mathbb{Z}} \sum_{\ell=1}^{L} a_\ell[n]\, h(t-t_\ell-nT), \tag{25}$$

where $\lambda = \{t_\ell\}_{\ell=1}^{L}$ is a set of unknown time delays contained in the time interval $[0, T]$, $\{a_\ell[n]\}$ are arbitrary bounded energy sequences and $h(t)$ is a known pulse shape. For a given set of delays λ, each signal of the form (25) lies in a shift-invariant subspace $\mathcal{A}_\lambda$, spanned by L generators $\{h(t-t_\ell)\}_{\ell=1}^{L}$. Since the delays can take on any values in the continuous interval $[0, T]$, the set of all signals of the form (25) constitutes an infinite union of shift-invariant subspaces $|\Lambda| = \infty$. Additionally, since any signal has parameters $\{a_\ell[n]\}_{n\in\mathbb{Z}}$, each of the $\mathcal{A}_\lambda$ subspaces has infinite cardinality, i.e., union type $\infty - \infty$. This model can represent, for example, a time-division multiple access (TDMA) setup, in which L transmitters access a joint channel on predefined time slots. Due to unknown propagations in the channel, the receiver intercepts symbol streams $a_\ell[n]$ at unknown delays t_ℓ.

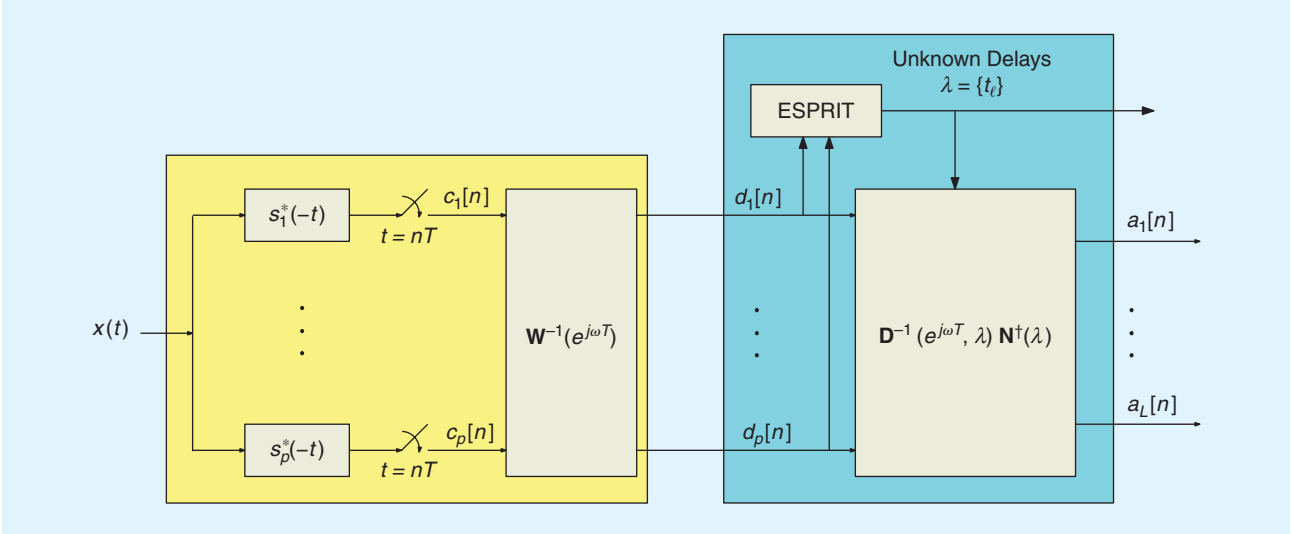
A sampling and reconstruction scheme for signals of the form (25) is depicted in Figure 15 [40]. The multichannel scheme has p parallel sampling channels. In each channel, the input signal $x(t)$ is filtered by a band-limited sampling kernel $s_\ell^*(-t)$ with frequency support contained in an interval of width $2\pi p/T$, followed by a uniform sampler at rate $1/T$, thus providing the sampling sequence $c_\ell[n]$. Note that just as in the MWC system, the multiple branches can be collapsed to a single filter whose output is sampled p times faster. The role of the sampling kernels is to smear the pulse in time, prior to low rate sampling.

To recover the signal from the samples, a properly designed digital filter correction bank, whose frequency-domain response is given by $\mathbf{W}^{-1}(e^{j\omega T})$, is applied to the sampling sequences $c_\ell[n]$. The entries of $\mathbf{W}(e^{j\omega T})$ depend on the choice of the sampling kernels $s_\ell^*(-t)$ and pulse shape $h(t)$ by

$$\mathbf{W}(e^{j\omega T})_{\ell,m} = \frac{1}{T} S_\ell^*(\omega + 2\pi m/T) H(\omega + 2\pi m/T). \tag{26}$$

The corrected sample vector $\mathbf{d}[n] = [d_1[n], \ldots, d_p[n]]^T$ is related to the unknown amplitude vector $\mathbf{a}[n] = [a_1[n], \ldots, a_L[n]]^T$ by a Vandermonde matrix that depends on the unknown delays t_ℓ [40]. Therefore, subspace detection methods, such as the estimation of signal parameters via rotational invariance techniques (ESPRIT) algorithm [72], can be used to recover the delays $\lambda = \{t_1, \ldots, t_L\}$. Once the delays are determined, additional filtering operations are applied on the samples to recover the information sequences $a_\ell[n]$. In particular, referring to Figure 15, the matrix $\mathbf{D}$ is a diagonal matrix with diagonal elements equal to $e^{-j\omega t_k}$, and $\mathbf{N}(\lambda)$ is a Vandermonde matrix with elements $e^{-j2\pi m t_k/T}$.

In general, the number of sampling channels p required to ensure unique recovery of the delays and sequences using the proposed scheme has to satisfy $p \geq 2L$ [40]. This leads



[FIG15] **Sampling and reconstruction scheme for signals of the form (25).**

to a minimal sampling rate of $2L/T$. For certain signals, the sampling rate can be reduced even further to $(L+1)/T$ [40]. Evidently, the minimal sampling rate is not related to the Nyquist rate of the pulse $h(t)$. Therefore, for wideband pulse shapes, the reduction in rate can be quite substantial. As an example, consider the setup in [73], used for characterization of ultrawideband wireless indoor channels. Under this setup, pulses with bandwidth of $W = 1$ GHz are transmitted at a rate of $1/T = 2$ MHz. Assuming that there are ten significant multipath components, this method reduces the sampling rate down to 40 MHz compared with the 2 GHz Nyquist rate.

NOISE-FREE VERSUS NOISY FRI MODELS

The performance of FRI techniques was studied in the literature mainly for noise-free cases. When the continuous-time signal $x(t)$ is contaminated by noise, recovery of the exact signal is no longer possible regardless of the sampling rate. Instead, one may speak of the minimum squared error (MSE) in estimating $x(t)$ from its noisy samples. In this case the rate of innovation L takes on a new meaning as the ratio between the best MSE achievable by any unbiased estimator and the noise variance σ^2, regardless of the sampling method [74]. This stands in contrast to the noise-free interpretation of L as the minimum sampling rate required for perfect recovery.

In general, the sampling rate that is needed to achieve an MSE of $L\sigma^2$ is equal to the rate associated with the affine hull Σ of the union set [74]. In some cases, this rate is finite, e.g., in a multiband union, but in many cases the sum covers the entire space $L_2(\mathbb{R})$, e.g., an FRI union with nonbandlimited pulse shape $h(t)$, in which case no finite-rate technique achieves the optimal MSE. This again is quite different from the noise-free case, in which recovery is usually possible at a rate of $2L$, where L is the individual subspace dimension.

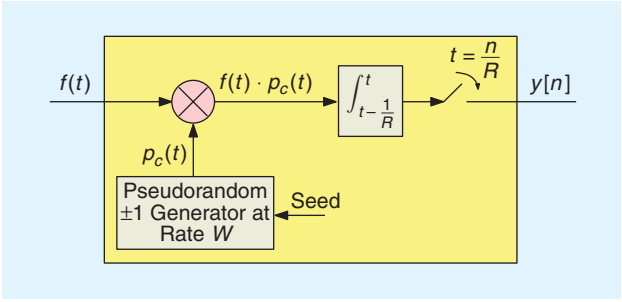
A consequence of these results is that oversampling can generally improve estimation performance. Indeed, it can be shown that sampling rates much higher than L are required in certain settings to approach the optimal performance. Furthermore, these gains can be substantial: In some cases, oversampling can improve the MSE by several orders of magnitude. These results help explain effects of numerical instability which are sometimes observed in FRI reconstruction. As a rule of thumb, it appears that for union of subspace signals, performance is improved at low rates if most of the unknown parameters identify the position within the subspace $\mathcal{A}_\lambda$, rather than the subspace index λ^*. Further details on these bounds and recovery performance appear in [74].

SPARSE SUM OF HARMONIC SINUSOIDS

DISCRETIZED MODEL

Rapidly growing interest in CS over the last few years has given a major drive to sub-Nyquist sampling. CS focuses on efficiently measuring a discrete signal (vector) $\mathbf{z}$ of length M that has $k < M$ nonzero entries. A measurement vector $\mathbf{y}$ of shorter length, proportional to k, is generated by $\mathbf{y} = \mathbf{\Phi z}$, using an underdetermined matrix $\mathbf{\Phi}$. Since $\mathbf{z}$ is sparse, it can be recovered from $\mathbf{y}$, even though $\mathbf{\Phi}$ has fewer rows than columns. "Sparse Solutions of Underdetermined Linear Systems" elaborates more on the techniques used in CS for sparse vector reconstruction. The CS setup borrows the sub-Nyquist terminology for the finite setting, so as to emphasize that the measurement vector $\mathbf{y}$ is shorter than $\mathbf{z}$.

Although CS is in essence a mathematical theory for measuring finite-length vectors, various researchers applied these ideas to sensing of analog signals by using discretized or finite-dimensional models [41], [75]–[77]. One of the first works in this direction [41] explores CS techniques for sensing a sparse sum of harmonic tones



[FIG16] Block diagram of the random demodulator [41]. (Figure courtesy of the IEEE.)

$$f(t) = \sum_{k=-(W/2-1)}^{W/2} a_k e^{j2\pi kt}, \quad \text{for } t \in [0,1) \tag{27}$$

with at most k nonzero coefficients a_k out of W possible tones. In contrast to FRI models that permit t_ℓ to lie on the continuum, the active sinusoids in (27) lie on a uniform grid of frequencies $\{k\Delta\}$ with normalized spacing $\Delta = 1$ (union type F − F).

The random demodulator (RD) senses a sparse harmonic input $f(t)$ by mapping blocks of Nyquist-rate samples to low rate measurements via a binary random combination, as depicted in Figure 16. The signal $f(t)$ is multiplied by a pseudorandom ± 1 generator alternating at rate W, and then integrated and dumped at a constant rate $R < W$. A vector $\mathbf{y}$ collects R consecutive measurements, resulting in the underdetermined system [41]

$$\mathbf{y} = \mathbf{\Phi f} = \mathbf{\Phi}\,\mathrm{DFT}\{\mathbf{z}\}, \tag{28}$$

where the random sign combinations are the entries of $\mathbf{\Phi}$ and $\mathbf{f}$ corresponds to the values of $f(t)$ on the Nyquist grid (more precisely, the entries of $\mathbf{f}$ are the values that are obtained by integrating-and-dumping $f(t)$ on $1/W$ time intervals). The vector of DFT coefficients $\mathbf{z}$ coincides with a_k due to the time-axis normalization $\Delta = 1$. Using CS recovery algorithms, $\mathbf{z}$ is determined and then $f(t)$ is resynthesized using (27). A bank of RD channels with overlapping integrations was studied in [48].

The RD method is one of the pioneer and innovative attempts to extend CS to analog signals. Underlying this approach is input modeling that relies on finite discretization. Thus, as long as the signal obeys this finite model, as in the case, for example, with harmonic tones (27), extending CS is possible following this strategy. In practice, however, in many applications we are interested in processing and representing an underlying analog signal, which is decidedly not finite-dimensional, e.g., multiband or FRI inputs. Applying discretization on analog signals that posses infinite structures can result in large hardware and software complexities, as we discuss in the next subsection.

DISCRETIZATION VERSUS CONTINUOUS ANALOG MODELING

Transition from analog to digital is one of the tricky parts in any sampling strategy. The approach we have been describing in this review treats analog signals by taking advantage of UoS modeling, where infiniteness enters either through the dimensions of the underlying subspaces $\mathcal{A}_\lambda$, the cardinality of the union $|\Lambda|$, or both (types F−∞, ∞−F and ∞−∞, respectively). The sparse harmonic model is, however, exceptional since in this case both Λ and $\mathcal{A}_\lambda$ are finite (type F−F). It is naturally tempting to use finite tools and to avoid the difficulties that come with infinite structures. Theoretically, an analog multiband signal can be approximated to a desired precision by a dense grid of discrete tones [41]. However, there is a practical

[TABLE 2] IMPACT OF DISCRETIZATION ON COMPUTATIONAL LOADS.

		RD			MWC
	DISCRETIZATION SPACING	**Δ = 1 HZ**	**Δ = 100 HZ**		
MODEL	K TONES	$300 \cdot 10^6$	$3 \cdot 10^6$	N BANDS	6
	OUT OF Q TONES	$10 \cdot 10^{10}$	$10 \cdot 10^8$	WIDTH B	50 MHZ
	ALTERNATION SPEED W	10 GHZ	10 GHZ	m CHANNELS	35
SAMPLING SETUP				M FOURIER COEFFICIENTS	195
				f_s PER CHANNEL	51 MHZ
	RATE R	2.9 GHZ	2.9 GHZ	TOTAL RATE	1.8 GHZ
PREPARATION					
COLLECT SAMPLES	NUM. OF SAMPLES N_R	$2.9 \cdot 10^9$	$2.9 \cdot 10^7$	$2N$ SNAPSHOTS OF $y[n]$	$12 \cdot 35 = 420$
DELAY	N_R/R	1 S	10 MS	$2N/f_s$	235NSEC
CS BLOCK					
MATRIX DIMENSIONS	$\mathbf{\Phi F} = N_R \times Q$	$2.6 \cdot 10^9 \times 10^{10}$	$2.6 \cdot 10^7 \times 10^8$	$\mathbf{C} = m \times M$	$\mathbf{35 \times 195}$
APPLY MATRIX	$\mathcal{O}(W \log W)$			$\mathcal{O}(mM + M\,logM)$	
STORAGE	$\mathcal{O}(W)$			$\mathcal{O}(mM)$	
REAL TIME (FIXED SUPPORT)					
MEMORY LENGTH	N_R	$2.9 \cdot 10^9$	$2.9 \cdot 10^7$	1 SNAPSHOT OF $\mathbf{y}[n]$	35
DELAY	N_R/R	1 S	10 MS	$1/f_s$	19.5 NS
MULT.-OPS. (MILLIONS/S)	$KN_R/(N_R/R)$	$8.7 \cdot 10^{11}$	$8.7 \cdot 10^9$	$2Nmf_s$	21420

price to pay—the finite dimensions grow arbitrarily large; a 1 MHz Nyquist-rate input boils down to a sparse recovery problem with $W = 10^6$ entries in $\mathbf{z}$. In addition, discretization brings forth sensitivity issues and loss of signal resolution as demonstrated in the sequel.

THE RD METHOD IS ONE OF THE PIONEER AND INNOVATIVE ATTEMPTS TO EXTEND CS TO ANALOG SIGNALS. UNDERLYING THIS APPROACH IS INPUT MODELING THAT RELIES ON FINITE DISCRETIZATION.

To highlight the issues that result from discretization of general analog models, we consider an example scenario of a wideband signal with $f_{NYQ} = 10$ GHz, 3 concurrent transmissions and 50 MHz bandwidths around unknown carriers f_i. Table 2, quoted from [51], compares various digital complexities of the MWC and an RD system which is applied on a Δ-spaced grid of frequencies for two discretization options $\Delta = 1$ Hz and $\Delta = 100$ Hz. The notation in the table is self-explanatory. It shows that discretization of general continuous inputs results in computational loads that effectively scale with the Nyquist rate of the input, which can sometimes be orders of magnitude higher than the complexity of approaches that directly treat the infinite union structure.

The differences reported in Table 2 stem from attempting to approximate a multiband model by a discrete set of tones, so as to consider inputs with comparable Nyquist bandwidth. At first sight, the signal models and compression techniques used in the MWC and RD seem similar, at least visually. A comprehensive study in [51] examines each system with its own model and compares them in terms of hardware and software complexities and robustness to model mismatches (as also briefly discussed in "Numerical Simulations of Sub-Nyquist Systems"). This comparison reveals that in this setting the MWC outperforms the RD, at least in these practical metrics, with a sampler hardware that can be readily implemented with existing analog devices and computationally light software algorithms. Similar conclusions were reached in [67], where sub-Nyquist radar imaging developed based on union modeling was demonstrated to accomplish accurate target identification and super-resolution capabilities in high signal-to-noise ratio (SNR) environments. In comparison, discretization-based approaches for radar imaging in high SNR were shown to suffer from spectral leakage which degrades identification accuracy and has limited super-resolution capabilities even in noise free settings.

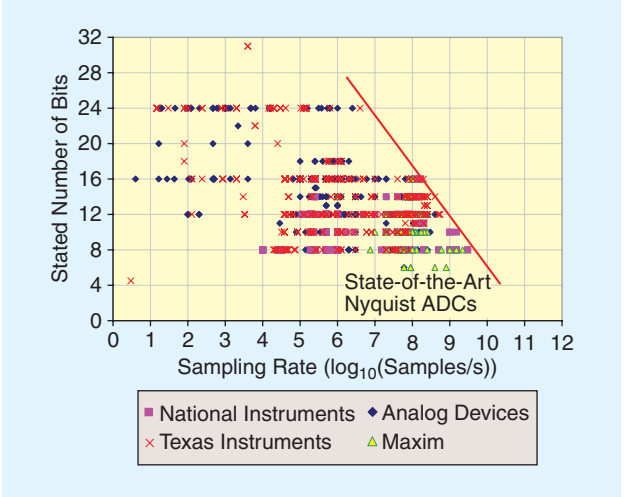
The conclusion we would like to convey is that union modeling provides a convenient mechanism to preserve the inherent infiniteness that many classes of analog signals posses. The infiniteness can enter thru the dimensions of the individual subspaces $\mathcal{A}_\lambda$, the union cardinality $|\Lambda|$, or both. Alternative routes relying on finite models to approximate continuous signals, presumably via discretization, may lead to high computational complexities and strong sensitivities. Nonetheless, there are examples of continuous-time signals that naturally possess finite representations (one such example is trigonometric polynomials). In such situations of an input that is well approximated by a regularized finite model of small size, analog discretization can be beneficial. It is therefore instructive to examine the specific acquisition problem at hand and choose between analog-based sampling to the discretization-based alternative. In either option, applying CS techniques in the digital domain, as part of reconstruction, can bring forward prominent advantages, i.e., provable robustness to noise and widely available off-the-shelf solvers. One potential application of CS is in the context of FRI recovery, where instead of using ESPRIT, MUSIC or annihilating filter for time-delay estimation on the continuum, one can consider discretizing the reconstruction time-axis and using a CS solver to increase the overall noise robustness [78].

In the next section, we summarize our review and discuss the potential of sub-Nyquist strategies to appear in real-world applications.

SUMMARY

FROM THEORY TO PRACTICE

We began the review with the Shannon-Nyquist theorem. Undoubtedly, uniform sampling ADC devices are the most common technology in the market. Figure 17 maps off-the-shelf ADC devices according to their sampling rate. The ADC industry has perpetually followed the Nyquist paradigm—the datasheets of all the devices that are reported in the figure highlight the conversion speed, referring to uniform sampling of the input. The industry is continuously striving to increase the possible uniform conversion rates.



[FIG17] ADC technology: Stated number of bits versus sampling rate. A map of more than 1,200 ADC devices from four leading manufacturers, according to online datasheets [47]. Previous mappings from the last decade are reported in [7] and [8].

NUMERICAL SIMULATIONS OF SUB-NYQUIST SYSTEMS

In this article, we have focused on bridging theory and practice, specifically on a high-level survey and comparison of sub-Nyquist methods. Such a high-level evaluation reveals the potential performance and inherent limitations in a device-independent setting. Numerical simulations are often used for these evaluation purposes. This box highlights delicate points concerning simulation of sub-Nyquist sampling strategies.

Hardware Modeling

A first step to numerical evaluation of an analog prototype is properly modeling the hardware components in a discrete computerized setup. For example, an analog filter can be represented by a digital filter with appropriate translation of absolute to angular frequencies. Modeling of an ADC device is a bit trickier. In classic PNS works [22], [23], [43], [45], the ADC is modeled as an ideal pointwise sampler. However, as explained in "Nyquist and Undersampling ADC Devices," a commercial ADC has an analog bandwidth limitation that dictates the maximal frequency b that the internal T/H circuitry can handle. To express the T/H limitations of the hardware, a lowpass filter preceding the pointwise sampling should be added [20]. Figure S4(a) depicts the spectrum of a single branch in a PNS setup. When discarding the analog bandwidth b, contents from high frequencies alias to baseband. Unfortunately, this result is misleading, since, in practice, the T/H bandwidth would eliminate the desired aliasing effect. This behavior is immediately noticed when inserting a lowpass filter with cutoff b before the ideal sampler.

Point Density

Once the hardware is properly modeled, the simulation computes samples on a grid of time points. The step size (in time) controls the accuracy of the computed samples compared with those that would have been obtained by the hardware. Clearly, if all the hardware nodes are bandlimited, then the computations can be performed at the Nyquist rate. The ADC is then visualized as a decimator at the end of the path. This option cannot be used, however, when the hardware nodes are not bandlimited. For example, in the MWC strategy, the periodic waveforms $p_i(t)$ are not necessarily bandlimited, and neither is the product $x(t)p_i(t)$, which, theoretically, consists of infinitely many shifts of the spectrum of $x(t)$. As a result the subsequent analog filtering, which involves continuous convolution between $h(t)$ and the nonbandlimited signal $x(t)p_i(t)$, becomes difficult to approximate numerically. In the simulations of [20], a simulation grid with density ten times higher than the Nyquist rate of $x(t)$ was used to estimate the MWC samples in a precision approaching that of the hardware. The figure below exemplifies the importance of correct density choice for simulation. The Fourier-series coefficients c_{il} of a sign alternating $p_i(t)$ were computed over a grid that contains r samples per each sign interval. Evidently, as r increases the simulation density improves and the coefficients converge to their true theoretical values.

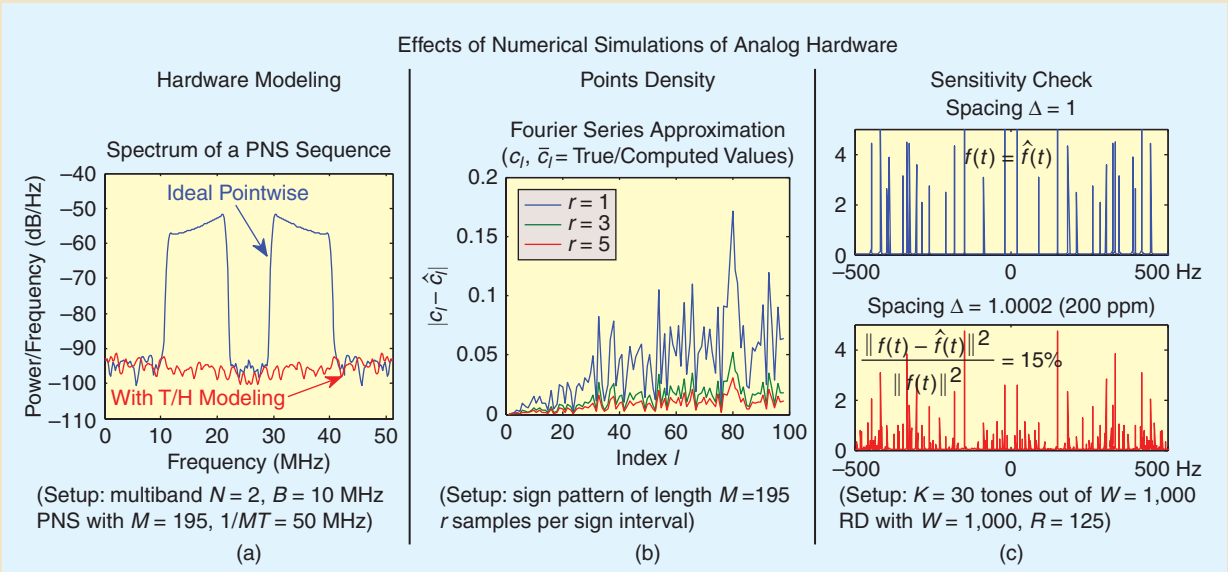
Sensitivity Check

Hardware circuits are prone to design imperfections. Therefore, besides simulation at the nominal working point, it is important to check the system behavior at nearby conditions; recall the wish list of Table 1. Figure S4(c) demonstrates the consequence of applying the RD on a harmonic sparse input, whose tones spacing Δ does not match exactly the spacing that the system was designed for. The reconstruction error is large; see also [51] and [84]. Numerical

Concluding this review, we would like to focus on multiband inputs and sketch the scenarios that may justify employing a sub-Nyquist solution over the traditional DSP scheme of Figure 1. Tables 3 and 4 summarize the sub-Nyquist methods we surveyed earlier. Among the subspace methods demodulation is already adapted by industry for sampling a multiband

[TABLE 3] SUB-NYQUIST STRATEGIES (SPECTRALLY SPARSE).

	STRATEGY	MODEL	CARDINALITY $\lvert\Lambda\rvert$	CARDINALITY $\mathcal{A}_\lambda$	ANALOG PREPROCESS.	REQ. ADC BANDWIDTH	RECONSTRUCTION PRINCIPLE	SUB-NYQ. PROCESS.	STATUS	TECHNOLOGY BARRIER
CLASSIC	SHANNON-NYQUIST	BANDLIMITED	1	∞		NYQUIST	INTERPOLATION (1)		COMMERCIAL	ADC
	DEMODULATION	MULTIBAND	1	∞	I/Q DEMOD.	LOW RATE	DAC + MODULATION	✓	COMMERCIAL	RF
	UNDERSAMPLING [18]	BANDPASS	1	∞	DELAY	NYQUIST	PIECEWISE FILTERING			ADC (T/H)
	PNS [22],[23], [43],[45]	MULTIBAND	1	∞	DELAY	NYQUIST	PIECEWISE FILTERING			ADC (T/H)
	PNS [42]	MULTIBAND	M	∞	DELAY	NYQUIST	CTF			ADC (T/H)
	FILTER-BANK [39]	SPARSE-SI	M	∞	FILTERS	NYQUIST	CTF			ADC (T/H)
UNION OF SUBSPACES	MWC [20]	MULTIBAND	M	∞	PERIODIC-MIXING	LOW	CTF	✓	2.2 GHZ BOARD-LEVEL [47]	RF
	RD [41]	SPARSE HARMONIC	K	W	RANDOM-SIGN MIXING	LOW	CS		800 KHZ BOARD-LEVEL [49]	SOFTWARE



[FIG S4] Accurate (red lines) versus incomplete (blue lines) numerical simulation of analog sub-Nyquist samplers: (a) PNS, (b) MWC, and (c) RD.

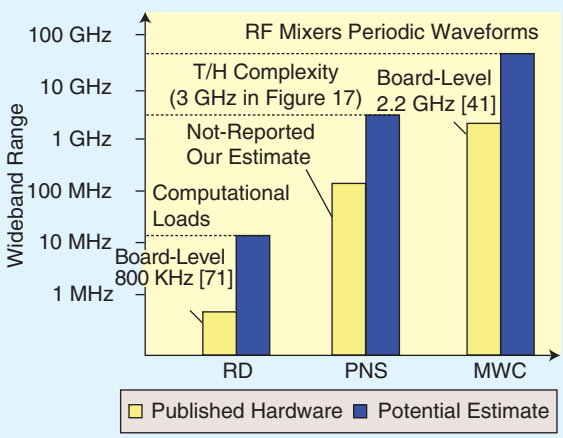
simulations in [20] and [50] and hardware experiments in [47] affirm robustness of the MWC system to various noise and imperfection sources.

We note that discretization techniques that are used to conduct a numerical study are not to be confused with model discretization approaches which use finite signal models to begin with. As an evaluation tool, discretization gives the ability to simulate the hardware performance to desired precision. Obviously, increasing the simulation density improves accuracy, at the expense of additional computations and memory and time resources. In practice, the hardware performs analog operations instantly, regardless of the run time and computational loads that were required for numerical simulations. In contrast, when basing the approach on discretization of the analog model, the choice of grid density brings forth issues of accuracy and various complexities to the actual sampling system. Eventually, model discretization also affects the size of problems that can be simulated numerically; multiband with 10 GHz Nyquist-rate in [20] versus a bandwidth of 32 kHz in [41].

input below f_{NYQ} when the carrier positions are known. Undersampling is also popular to some extent when there is a single band of information, and the maximal frequency $f_u < b$, namely within the available T/H bandwidths of commercial ADC devices. In contrast, although popular in time-interleaved ADCs, PNS was not widely embraced for sub-Nyquist sampling.

[TABLE 4] SUB-NYQUIST STRATEGIES (FINITE RATE OF INNOVATION).

	CARDINALITY		RECONSTRUCTION
ANALOG PREPROCESSING	$\lvert\Lambda\rvert$	$\mathcal{A}_\lambda$	ALGORITHM
LOWPASS [27], [62]	∞	$2L$	ANNIHILATING FILTER
GAUSSIAN [27], [62]	∞	$2L$	ANNIHILATING FILTER
POLY.-/EXP.-REPRODUCING KERNEL [28]	∞	$2L$	MOMENTS FILTERING
SUCC.-INTEGRATION [69]	∞	$2L$	ANNIHILATING FILTER
EXP.-FILTERING [70]	∞	$2L$	POLE-CANCELATION FILTER
RC-CIRCUIT [71]	∞	2	CLOSED-FORM
SOS-FILTERING [63]	∞	$2L$	ANNIHILATING FILTER
PERIODIC-MIXING [64]	∞	$2L/\infty$	ANNIHILATING FILTER
FILTER-BANK [40]	∞	∞	MUSIC [79] / ESPRIT [72]



[FIG18] Technology potential of state-of-the-art sub-Nyquist strategies (for multiband inputs).

This situation is perhaps reasoned by the fact that the technology barrier of any pointwise method is eventually limited by the analog bandwidth of the T/H stage. Accumulating wideband noise is another drawback of PNS (and the MWC and RD).

When the carrier frequencies are unknown, single subspace methods are not an option anymore. In Figure 18, we draw the rough potential of three leading sub-Nyquist technologies (for multiband inputs) as we foresee. The MWC approach extends the capabilities of I/Q demodulation by mixing the input with multiple sinusoids at once, probably limited by noise. T/H limitations remain the bottleneck of PNS alternatives [42]. We added the RD to Figure 18 despite the fact that it treats harmonic inputs rather than narrowband transmissions. The figure reveals that while hardware constraints bound the potential of sampling strategies such as MWC and PNS, it is the software complexity that limits the RD approach, since complexity of its recovery algorithm scales with the high Nyquist rate W.

The MWC provides a sampling solution for scenarios in which the input reaches frequencies that are beyond the analog bandwidths of commercial ADCs, $f_{\max} >$ b. The system can be used when knowledge of the carrier positions is present or absent. Furthermore, even when $f_{\max}$ is moderate, say within T/H bandwidth b of available ADC devices, the MWC proposes an advantage of reducing the processing rates, so that a cheap processor can be used instead of a premium device that can accommodate the Nyquist rate. In fact, even when the sole purpose of the system is to store the samples, the cost of storage devices that are capable of handling high-speed bursts of data streams, with or without compression, may be a sufficient motivation to shift from Figure 1 toward the MWC sub-Nyquist system. The hardware prototype [47] is also applicable for sampling FRI signals at their innovation rate [66]. A recent publication [51] introduces Xampling, a generic framework for signal acquisition and processing in UoS. The Xampling paradigm is built upon the various insights and example applications we surveyed in this review and the general sampling approach developed in [39].

Finally, we would like to point out the Nyquist-folding receiver of [80] as an alternative sub-Nyquist paradigm. This method proposes an interesting route to sub-Nyquist sampling. It introduces a deliberate jitter in an undersampling grid, which induces a phase modulation at baseband such that the modulation magnitude depends on the unknown carrier position. We did not elaborate on [80] since a reconstruction algorithm was not published yet. In addition, the jittered sampling results in a time-varying operator and thus departs from the linear time-invariant framework that unifies all the works we surveyed herein. However, this is an interesting venue for developing sub-Nyquist strategies.

FORECAST: SUB-NYQUIST IN COGNITIVE RADIOS

Sub-Nyquist systems may play an important role in the next generation of communication systems. The traditional zero-IF and low-IF receivers are based on demodulation by a given carrier frequency f_c prior to sampling. Knowledge of the carrier frequency is utilized to improve circuit properties of the receiver for the given f_c, at the expense of degraded performance in spectrum zones that are far from the specified frequency. For example, the oscillator that generates f_c in the I/Q-demodulator can be chosen to have a narrow tuning range so as to improve the frequency stability. An active mixer whose linear range is tailored to f_c is another possible design choice once the carrier is known.

In the last decade, the trend is to construct generic hardware platforms to reduce the production expenses involved in specifying the design for a given carrier. Two strategies that are recently being pushed forward are:

- *software-defined radio (SDR)* [81], where the receiver contains a versatile wideband hardware platform. The firmware is programmed to a specific f_c after manufacturing, enabling the SDR to function in different countries or by several cellular operators.
- *cognitive radio (CR)* [82], which adds another layer of programming, by permitting the software to adjust the working frequency f_c according to high-level cognitive decisions, such as cost of transmission and availability of frequency channels.

The interest in CR devices stems from an acute shortage in additional frequency regions for licensing, due to past allocation policies of spectral resources. Fortunately, studies have shown that those licensed regions are not occupied most of the time. The prime goal of a CR device is to identify these unused frequency regions and utilize them while their primary user is idle. Today, most civilian applications assume knowledge of carrier frequencies so that standard demodulation is possible. In contrast, CR is an application where by definition spectral support varies and is unknown a priori. We therefore foresee sub-Nyquist sampling playing an important role in future CR platforms. The MWC hardware, for instance, does not assume the carrier positions and is therefore designed in a generic way to cover a wideband range of frequencies. The ability to recover the frequency support from lowrate sampling may be the key to efficient spectrum sensing in CR [83].

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments and insightful suggestions that helped improve the manuscript. Moshe Mishali is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities. Yonina C. Eldar's work was supported in part by the Israel Science Foundation under grant 170/10 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless Communications NEWCOM++ (contract 216715).

AUTHORS

Moshe Mishali (moshiko@tx.technio.ac.il) received the B.Sc. degree (summa cum laude) in 2000 and the Ph.D. degree in 2011, both in electrical engineering in 2000, from the Technion–Israel Institute of Technology. From 1996 to

2000, he was a member of the Technion Program for Exceptionally Gifted Students. Since 2006, he has been a research assistant and project supervisor with the Signal and Image Processing Lab, Electrical Engineering Department, Technion. His research interests include theoretical aspects of signal processing, compressed sensing, sampling theory, and information theory. He received the 2008 Hershel Rich Innovation Award.

Yonina C. Eldar (yonina@ee.technion.ac.il) received the B.Sc. degree in physics and B.Sc. degree in electrical engineering both from Tel-Aviv University (TAU), Israel, in 1995 and 1996, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2002. She is currently a professor in the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa. She is also a research affiliate with the Research Laboratory of Electronics at MIT and a visiting professor at Stanford University. She was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. Her list of awards includes the 2004 Wolf Foundation Krill Prize for Excellence in Scientific Research; the 2005 Andre and Bella Meyer Lectureship; the 2007 Henry Taub Prize for Excellence in Research; the 2008 Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award; the 2009 Technion's Award for Excellence in Teaching; the 2010 Michael Bruno Memorial Award from the Rothschild Foundation; and the 2011 Weizmann Prize for Exact Sciences. She is a Senior Member of the IEEE.

REFERENCES

[1] A. J. Jerri, "The shannon sampling theorem—Its various extensions and applications: A tutorial review," *Proc. IEEE*, vol. 65, no. 11, pp. 1565–1596, 1977.

[2] P. L. Butzer, "A survey of the Whittaker–Shannon sampling theorem and some of its extensions," *J. Math. Res. Exposition*, vol. 3, no. 1, pp. 185–212, 1983.

[3] M. Unser, "Sampling—50 years after Shannon," *Proc. IEEE*, vol. 88, no. 4, pp. 569–587, Apr. 2000.

[4] P. P. Vaidyanathan, "Generalizations of the sampling theorem: Seven decades after Nyquist," *IEEE Trans. Circuits Syst. I*, vol. 48, no. 9, pp. 1094–1109, Sept. 2001.

[5] A. Aldroubi and K. Gröchenig, "Non-uniform sampling and reconstruction in shift-invariant spaces," *SIAM Rev.*, vol. 43, no. 4, pp. 585–620, Mar. 2001.

[6] Y. C. Eldar and T. Michaeli, "Beyond bandlimited sampling," *IEEE Signal Processing Mag.*, vol. 26, no. 3, pp. 48–68, May 2009.

[7] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE J. Select. Areas Commun.*, vol. 17, no. 4, pp. 539–550, 1999.

[8] L. Bin, T. W. Rondeau, J. H. Reed, and C. W. Bostian, "Analog-to-digital converters," *IEEE Signal Processing Mag.*, vol. 22, no. 6, pp. 69–77, Nov. 2005.

[9] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, pp. 10–21, Jan. 1949.

[10] H. Nyquist, "Certain topics in telegraph transmission theory," *Trans. AIEE*, vol. 47, no. 2, pp. 617–644, Apr. 1928.

[11] E. T. Whittaker, "On the functions which are represented by the expansion of interpolating theory," *Proc. R. Soc. Edinb.*, vol. 35, pp. 181–194, 1915.

[12] V. A. Kotelńikov, "On the transmission capacity of "ether" and wire in electrocommunications," *Izd. Red. Upr. Svyazi RKKA (Moscow)*, 1933.

[13] A. Papoulis, "Error analysis in sampling theory," *Proc. IEEE*, vol. 54, no. 7, pp. 947–955, July 1966.

[14] A. Papoulis, "Generalized sampling expansion," *IEEE Trans. Circuits Syst.*, vol. 24, no. 11, pp. 652–654, Nov. 1977.

[15] J. Crols and M. S. J. Steyaert, "Low-IF topologies for high-performance analog front ends of fully integrated receivers," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 3, pp. 269–282, Mar. 1998.

[16] N. Boutin and H. Kallel, "An arctangent type wideband PM/FM demodulator with improved performances," in *Proc. 33rd Midwest Symp. Circuits and Systems*, 1990, pp. 460–463.

[17] J. G. Proakis and M. Salehi, *Digital Communications*. New York: McGraw-Hill, 1995.

[18] R. G. Vaughan, N. L. Scott, and D. R. White, "The theory of bandpass sampling," *IEEE Trans. Signal Processing*, vol. 39, no. 9, pp. 1973–1984, Sept. 1991.

[19] H. J. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Math.*, vol. 117, pp. 37–52, Feb. 1967.

[20] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE J. Select. Topics Signal Processing*, vol. 4, no. 2, pp. 375–391, Apr. 2010.

[21] D. M. Akos, M. Stockmaster, J. B. Y. Tsui, and J. Caschera, "Direct bandpass sampling of multiple distinct RF signals," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 983–988, 1999.

[22] A. Kohlenberg, "Exact interpolation of band-limited functions," *J. Appl. Phys.*, vol. 24, pp. 1432–1435, Dec. 1953.

[23] Y.-P. Lin and P. P. Vaidyanathan, "Periodically nonuniform sampling of bandpass signals," *IEEE Trans. Circuits Syst. II*, vol. 45, no. 3, pp. 340–351, Mar. 1998.

[24] W. Black and D. Hodges, "Time interleaved converter arrays," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 1980, vol. XXIII, pp. 14–15.

[25] C. Vogel and H. Johansson, "Time-interleaved analog-to-digital converters: Status and future directions," in *Proc. 2006 Int. Symp., Circuits and Systems*, no. 4, pp. 3386–3389, 2006.

[26] P. Nikaeen and B. Murmann, "Digital compensation of dynamic acquisition errors at the front-end of high-performance A/D converters," *IEEE Trans. Signal Processing*, vol. 3, no. 3, pp. 499–508, June 2009.

[27] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1417–1428, 2002.

[28] P. L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang Fix," *IEEE Trans. Signal Processing*, vol. 55, no. 5, pp. 1741–1757, May 2007.

[29] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2334–2345, June 2008.

[30] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[31] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[32] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[33] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Signal Processing*, vol. 59, no. 9, pp. 4053–4085, Sept. 2011.

[34] M. Mishali and Y. C. Eldar, "Xampling: Compressed sensing of analog signals," *Compressed Sensing: Theory and Applications*. Y. C. Eldar and G. Kutyniok, Eds. Cambridge University Press, Cambridge, U.K., no. 3, 2012.

[35] Y. C. Eldar and T. G. Dvorkind, "A minimum squared-error framework for generalized sampling," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 2155–2167, June 2006.

[36] C. de Boor, R. A. DeVore, and A. Ron, "The structure of finitely generated shift-invariant spaces in $L_2(\mathbb{R}^d)$" *J. Funct. Anal.*, vol. 119, no. 1, pp. 37–78, 1994.

[37] O. Christensen and Y. C. Eldar, "Generalized shift-invariant systems and frames for subspaces," *J. Fourier Anal. Applicat.*, vol. 11, no. 3, pp. 299–313, 2005.

[38] M. Mishali and Y. C. Eldar, "Expected-RIP: Conditioning of the modulated wideband converter," in *Proc. IEEE Information Theory Workshop (ITW 2009)*. New York: IEEE, Oct. 2009, pp. 343–347.

[39] Y. C. Eldar, "Compressed sensing of analog signals in shift-invariant spaces," *IEEE Trans. Signal Processing*, vol. 57, no. 8, pp. 2986–2997, Aug. 2009.

[40] K. Gedalyahu and Y. C. Eldar, "Time delay estimation from low rate samples: A union of subspaces approach," *IEEE Trans. Signal Processing*, vol. 58, no. 6, pp. 3017–3031, June 2010.

[41] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.

[42] M. Mishali and Y. C. Eldar, "Blind multi-band signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Signal Processing*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[43] C. Herley and P. W. Wong, "Minimum rate sampling and reconstruction of signals with arbitrary frequency support," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1555–1564, July 1999.

[44] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," in *Proc. IEEE Int. Conf. ASSP*, May 1996, vol. 3, pp. 1688–1691.

[45] R. Venkataramani and Y. Bresler, "Perfect reconstruction formulas and bounds on aliasing error in sub-Nyquist nonuniform sampling of multiband signals," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2173–2183, Sept. 2000.

[46] M. Mishali and Y. C. Eldar, "Reduce and boost: Recovering arbitrary sets of jointly sparse vectors," *IEEE Trans. Signal Processing*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.

[47] M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan, "Xampling: Analog to digital at sub-Nyquist rates," *IET Circuits, Devices Syst.*, vol. 5, no. 1, pp. 8–20, Jan. 2011.

[48] Z. Yu, S. Hoyos, and B. M. Sadler, "Mixed-signal parallel compressed sensing and reception for cognitive radio," in *Proc. ICASSP*, 2008, pp. 3861–3864.

[49] T. Ragheb, J. N. Laska, H. Nejati, S. Kirolos, R. G. Baraniuk, and Y. Massoud, "A prototype hardware for random demodulation based compressive analog-to-digital conversion," in *Proc. 51st Midwest Symp. Circuits and Systems (MWSCAS 2008)*, pp. 37–40.

[50] Y. Chen, M. Mishali, Y. C. Eldar, and A. O. Hero III, "Modulated wideband converter with non-ideal lowpass filters," in *Proc. ICASSP*, 2010, pp. 3630–3633.

[51] M. Mishali, Y. C. Eldar, and A. Elron, "Xampling: Signal acquisition and processing in union of subspaces," submitted for publication.

[52] F. Gardner, "Properties of frequency difference detectors," *IEEE Trans. Commun.*, vol. 33, no. 2, pp. 131–138, Feb. 1985.

[53] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 2197–2202, Mar. 2003.

[54] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. Conf. Rec. 27th Asilomar Conf. Signals, Systems and Computers*, 1993, pp. 40–44.

[55] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.* (Special Issue on Sparse Approximations in Signal and Image Processing), vol. 86, pp. 572–588, Apr. 2006.

[56] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.* (Special Issue on Sparse Approximations in Signal and Image Processing), vol. 86, pp. 589–602, Apr. 2006.

[57] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Trans. Signal Processing*, vol. 54, no. 12, pp. 4634–4643, Dec. 2006.

[58] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Processing*, vol. 53, no. 7, pp. 2477–2488, July 2005.

[59] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[60] Y. C. Eldar and G. Kutyniok, Eds, *Compressed Sensing: Theory and Applications*, Cambridge University Press, Cambridge, U.K., 2012.

[61] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[62] I. Maravic and M. Vetterli, "Sampling and reconstruction of signals with finite rate of innovation in the presence of noise," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 2788–2805, 2005.

[63] R. Tur, Y. C. Eldar, and Z. Friedman, "Innovation rate sampling of pulse streams with application to ultrasound imaging," *IEEE Trans. Signal Processing*, vol. 59, no. 4, pp. 1827–1842, Apr. 2011.

[64] K. Gedalyahu, R. Tur, and Y. C. Eldar, "Multichannel sampling of pulse streams at the rate of innovation," *IEEE Trans. Signal Processing*, vol. 59, no. 4, pp. 1491–1504, Apr. 2011.

[65] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1997.

[66] M. Mishali, R. Hilgendorf, E. Shoshan, I. Rivkin, and Y. C. Eldar, "Generic sensing hardware and real-time reconstruction for structured analog signals," in *Proc. ISCAS*, 2011, pp. 1748–1751.

[67] W. U. Bajwa, K. Gedalyahu, and Y. C. Eldar, "Identification of parametric underspread linear systems and super-resolution radar," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2548–2561, June 2011.

[68] E. Matusiak and Y. C. Eldar. (2010, Oct.). Sub-Nyquist sampling of short pulses: Theory. *IEEE Trans. Inform. Theory* [Online]. Available: arXiv.org 1010.3132

[69] J. Kusuma and V. Goyal, "Multichannel sampling of parametric signals with a successive approximation property," in *Proc. IEEE Int. Conf. Image Procesing (ICIP)*, Oct. 2006, pp. 1265–1268.

[70] H. Olkkonen and J. T. Olkkonen, "Measurement and reconstruction of impulse train by parallel exponential filters," *IEEE Signal Processing Lett.*, vol. 15, pp. 241–244, 2008.

[71] C. Seelamantula and M. Unser, "A generalized sampling method for finite-rate-of-innovation-signal reconstruction," *IEEE Signal Processing Lett.*, vol. 15, pp. 813–816, 2008.

[72] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 7, pp. 984–995, July 1989.

[73] M. Z. Win and R. A. Scholtz, "Characterization of ultra-wide bandwidth wireless indoor channels: A communication-theoretic view," *IEEE J. Select. Areas Commun.*, vol. 20, no. 9, pp. 1613–1627, Dec. 2002.

[74] Z. Ben-Haim, T. Michaeli, and Y. C. Eldar. (2010, Sept.). Performance bounds and design criteria for estimating finite rate of innovation signals. *IEEE Trans. Inform. Theory* [Online]. Available: http://arxiv.org/pdf/1009.2221

[75] A. W. Habboosh, R. J. Vaccaro, and S. Kay, "An algorithm for detecting closely spaced delay/Doppler components," in *Proc. ICASSP*, Apr. 1997, pp. 535–538.

[76] W. U. Bajwa, A. M. Sayeed, and R. Nowak, "Learning sparse doubly-selective channels," in *Proc. Allerton Conf. Communication, Control, and Computing*, Sept. 2008, pp. 575–582.

[77] M. A. Herman and T. Strohmer, "High-resolution radar via compressed sensing," *IEEE Trans. Signal Processing*, vol. 57, no. 6, pp. 2275–2284, June 2009.

[78] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Processing*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[79] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. 34, no. 3, pp. 276–280, Mar. 1986. First presented at RADC Spectrum Estimation Workshop, Griffiss AFB, New York, 1979.

[80] G. L. Fudge, R. E. Bland, M. A. Chivers, S. Ravindran, J. Haupt, and P. E. Pace, "A Nyquist folding analog-to-information receiver," in *Proc. 42nd Asilomar Conf. Signals, Systems and Computers*, Oct. 2008, pp. 541–545.

[81] J. Mitola, "The software radio architecture," *IEEE Commun. Mag.*, vol. 33, no. 5, pp. 26–38, May 1995.

[82] J. Mitola, III, "Cognitive radio for flexible mobile multimedia communications," *Mobile Netw. Applicat.*, vol. 6, no. 5, pp. 435–441, 2001.

[83] M. Mishali and Y. C. Eldar, "Wideband spectrum sensing at sub-Nyquist rates," *IEEE Signal Processing Mag.*, vol. 28, no. 4, pp. 102–135, July 2011.

[84] M. F. Duarte and R. G. Baraniuk. (2010). Spectral compressive sensing. [Online]. Available: http://www.math.princeton.edu/~mduarte/images/SCS-TSP.pdf

[85] Y. Eldar. (2011). Xampling: Analog to digital at sub-Nyquist rates. [Online]. Available: http://webee.technion.ac.il/Sites/People/YoninaEldar/Info/hardware.html

[SP]

Harvey F. Silverman

One City—Two Giants: Armstrong and Sarnoff: Part 1

EDITOR'S INTRODUCTION

Our guest in this column is Dr. Harvey F. Silverman. Dr. Silverman received his Ph.D. and Sc.M. degrees from Brown University. He earned his B.S.E. and B.S. degrees from Trinity College in Hartford, Connecticut. He worked at IBM T.J. Watson Research Center in Yorktown Heights, New York, between 1970 and 1980. During these ten years, he worked on a number of projects related to image processing methods applied to earth resources satellite data, analytical methods for computer system performance, speech recognition, and the development of a real-time I/O system as the manager of the Speech Terminal Project. In 1980, he joined Brown University as a professor of engineering. From 1991 to 1998, he was the dean of engineering at Brown University. Dr. Silverman is a coauthor on more than 215 journal and conference papers. Currently, he conducts research on detection of autism from infant cries, microphone arrays, acoustics, and source-location estimation.

Dr. Silverman is an IEEE Life Fellow, was a member of the Executive Committee, Trustees at Trinity College from 2001 to 2003, and a Charter Trustee at Trinity from 1994 to 2003. He was nominated for the 1994 ComputerWorld Smithsonian Award and received the IEEE Centennial Medal Award in 1984. In 1981, he received the IEEE ASSP Society Meritorious Service Award, and he received a number of awards at IBM for his outstanding work on speech recognition and detection algorithms. At Brown, he supervised and graduated 26 doctoral students, and currently he supervises three Ph.D. students.

Dr. Silverman has been an avid vegetable gardener for over 40 years and currently has about 6,000 square feet under cultivation. An outgrowth of this passion is that of using his knowledge and engineering skills to develop intelligent means for deterring the many furry and feathered creatures who very much like to share in the eating, although not the planting and cultivating. So far, the critters are winning!

As our author mentions at the beginning of this article, in 1969 he bought the book *Man of High Fidelity*, which was a 1969 paperback version of a 1956 biography of Edwin Howard Armstrong. The book was so intriguing that Dr. Silverman borrowed an autobiography of David Sarnoff. This was the beginning of a journey where our author learned about these two inventors and how such two strong characters and complex men spent more than 40 years in competition and race to inventions and pioneering. Our author has made it a habit to educate his students at Brown on a weekly basis about history through Sarnoff and Armstrong and the dynamics of inventions including patents, trade secrets, competition, restraint of competition, and politics. Dr. Silverman suggests this story to be an important part of electrical engineering education. Of course, some people have a different view of the two men and their stories, and our author points this out in his article citing a 1956 letter written by Lee de Forest.

Dear readers, Dr. Silverman is taking us on an intriguing journey in the complex lives of Armstrong and Sarnoff. In this column, we publish Part 1 of the journey; in the next issue, we will publish the second and final part of this journey.

Ghassan AlRegib

As an avid reader, crazy-gadget-loving engineer, and history enthusiast, I bought a US$1 paperback when I was a graduate student in 1969 titled *Man of High Fidelity*, the 1969 paperback version of the 1956 biography of Edwin Howard Armstrong by Lawrence Lessing. The story told within was so fascinating that I needed to learn more, so I borrowed a copy (it had been given to an outstanding Radio Corporation of America (RCA) researcher by Sarnoff himself) of the commissioned autobiography of David Sarnoff written by his cousin, Eugene Lyons, published in 1966. It was clear that these two giants of their time had lives that intermingled over more than 40 years. I loaned out the paperback and never had it returned—aside from this, I did little to learn more until about 20 years ago, when I thought it would be a good idea for my sophomore electrical circuits students at Brown to hear their story, and thus of the beginning of the electronics era through the history of radio invention. Therefore, each Friday of a semester for about ten minutes each week, I have been telling the story I will try to relate here.

The story is poignant; different authors (and original participants) have had widely different viewpoints, as the men were strong characters and complex, to say the least.

Digital Object Identifier 10.1109/MSP.2011.942836
Date of publication: 1 November 2011

I shall try to tell it as someone quite inspired by many of the qualities of both men and who has hopefully learned a little by seeing some instances in which some compromising would have led to better outcomes.

I do point out another wonderful source—Ken Burns produced a powerful 2 h documentary called "Empire of the Air" about ten years ago. Also, would you believe that there is a new opinion on the story of FM radio expressed in a 2010 book by Gary L. Frost, *Early FM Radio*?

At a recent International Conference on Acoustics, Speech, and Signal Processing (ICASSP), it became clear that the same situations faced by Armstrong, Sarnoff, and others in the early days of the electronics industry still happen today. Trade secrets, patents, competition, restraint of competition, politics, and just simple fate all coalesce to determine the winners and losers in our vast electronics community just as they did back then. It is for this reason that I suggest that the story should be a part of an electrical engineer's education. With this article, I hope that a few more people will know some history and thus not be doomed to repeat it.

THE BEGINNING

The turn into the 20th century was indeed a time for beginnings. There were no "electronics," although the electrical age was well along. Morse's telegraph in 1839 had been the first practical invention that harnessed this new force—electricity. A few decades later, electric lighting followed, with electric motors finding great utility soon thereafter. This naturally allowed the development of generators and then power systems for cities. The telephone was an invention contemporary to electric lighting and used early forms of microphones that converted sound to an electronic analog that could be sent along wires—continuous analog signals instead of the (digital?) dots and dashes of the telegraph. However, there was no good way to amplify these electronic analogs. In the United States, large corporations such as General Electric (GE), Westinghouse, and American Telephone and Telegraph became the movers of this new electrical industry with new opportunities in motors, generators, power distribution, telephony, and lighting so large as to turn these electrical corporate giants into the growth industry of their day. GE and AT&T started the two largest research laboratories for this new industry, while Westinghouse made advances using AC ideas, based on advice and counsel of Nikola Tesla.

There were some physicists and engineers of this period who were fascinated by the possibilities of wireless transmission and electronic phenomena. Perhaps Heinrich Hertz is a good place to start when he, in 1888, demonstrated that wireless transmission would occur using a spark generator and a "tuned" C-shaped loop that sparked across its open end across a room when the spark was initiated by the generator. It took a few years, but this principle was demonstrated further by Guglielmo Marconi in 1899, transmitting telegraphic dots and dashes—a device that appeared to have some useful potential! Soon thereafter, the Marconi Company in England was founded with the American subsidiary, American Marconi, coming a few years later.

IT IS HARD TO BELIEVE THAT THE CAREER OF ONE OF THE MOST INFLUENTIAL MEN OF THE EARLY WIRELESS INDUSTRY WAS DETERMINED BY SUCH KARMA!

During these years too, some experiments of interest were carried out based on the light bulb. In 1873, Thomas Guthrie reported thermionic emission, but in 1884, Edison, ever the trial-and-error experimenter, added a second filament to his vacuum bulb and saw that electricity would flow from one filament to the other under certain circumstances. While Edison patented what he found (US patent 307031), and this was to be his only true contribution to "science," he was unable to see the value of the invention nor understand the underlying physics. This was dubbed the Edison effect and abandoned by the great inventor. Edison was not going to be a player in the electronics field.

In July 1900, two nine-year-olds of vastly different backgrounds were living in New York City. The older of the two, Edwin Howard Armstrong (18 December 1890) was born into a "genteel Victorian household" [7]. He was the son of John Armstrong, who worked for Oxford Press. John ultimately rose to the position of the manager of the American branch and was a trustee of the old North Presbyterian Church in Manhattan. His mother, Emily Smith Armstrong was a graduate of Hunter College and had been a public school teacher for ten years prior to marrying John Armstrong in 1888. Edwin was the oldest of three, having two younger sisters. At age nine, he contracted St. Vitus' Dance, an early childhood disease now associated with rheumatic fever. While St. Vitus' Dance normally ran its course in a few months, it was a full two years that Edwin remained at home, schooled by his family until his recovery. To get him "into the sun," [7] the family moved to Yonkers at this time to a large house with an attic in which he could do his experiments. While he fully recovered, he was left with a lifelong "tic" in which he would "hitch his shoulder forward and twist his neck and mouth whenever he was excited or under stress" [7]. In 1904 and 1905, his father, returning from annual trips to London, brought Edwin two new books, *The Boys' Book of Inventions* and *Stories of Inventions*. It was thus at age 14 that he decided to be an inventor—his path was set early.

The other nine-year old was David Sarnoff, who, at this time, had literally just stepped off a steerage deck, having emigrated from a shtetl (a small town) called Uzlian in the province of Minsk in Tsarist Russia. He had been born just two months after Armstrong on 27 February 1891, the oldest of three sons. By age five, David showed himself to be handsome and bright and knew how to read and to recite passages from the Bible by heart. In 1896, his father, Abraham Sarnoff, left the family for America to earn the money necessary to bring the family over to join

him. Abraham was pious but sickly, and it took him four years. During this period, David's mother was struggling in Uzlian, so his grandmother arranged for him to leave the family for four years to "become learned" [9]. He was sent 400 miles away to live and study with his great uncle, a rabbi in Korme. He was following the traditional path for the learned, studying 12–14 h a day reading over the Prophets in Hebrew and, later, the Talmud in Aramaic. Perhaps it was in this early period that he truly sharpened his mind, learning the disciplines of patience, concentration, and memory. After four years, his mother sent for him—they were going to America.

America! When David Sarnoff arrived in New York, he spoke no English and moved into a squalid, narrow, fourth-floor railroad flat of a decrepit tenement in the Lower East Side of Manhattan. Abraham Sarnoff's condition grew worse, and he could barely support his family. Ultimately, he died a few years later when David was a young teenager. However, the few years Abraham gave to his family allowed David to attend public school, originally being placed in one of the special classes set aside for new immigrants. He was reading English and speaking with some fluency before the end of the year. At the same time, he had to help support the family as the oldest son. He started by peddling Yiddish newspapers, rushing after school to grab a bundle, making a profit of US$0.25 if he could sell 50 papers in a very competitive selling environment. When his father died, David was no longer able to attend school, so he began working full time. About the time of his bar mitzvah at age 13, he had turned his newspaper peddling into a family newsstand at the corner of 46th Street and Tenth Avenue, the heart of a tough Irish neighborhood called Hell's Kitchen. As it was for several turning points in his life, good fortune played a role; the US$200 needed to purchase the newsstand was given him by an anonymous, mysterious middle-aged woman who came to his tenement! Twenty years later he found out she was a social worker, but he never found out who the real donor of the gift was. The tough Irish kids did not make life easy for Jewish boys in their neighborhood, but David held his own although forced by the circumstance to grow in toughness. He successfully moved his family to the area and a slightly better apartment so that they could all play roles in their new business.

Armstrong did not have the same finishing school of the streets. He went on to high school and college, was the captain of the Yonkers High School tennis team, finished his high school career with an 89.8 average, and was accepted into the Columbia University School of Engineering, Department of Electrical Engineering. During his high school days, he had become fascinated with the art of wireless telegraphy and telephony, had erected a large, 150 ft antenna in his backyard and was making many new acquaintances over the miniscule electrical currents in the air using the crude apparatus of the day. He entered college in 1909, commuting to school on his new Indian motorcycle his father had bought for him as a graduation present, and his first years were undistinguished. He had a one-track mind—wireless—and what he did not like he largely ignored, but passed his courses. He spent much time in his attic communicating with his amateur-radio cadre.

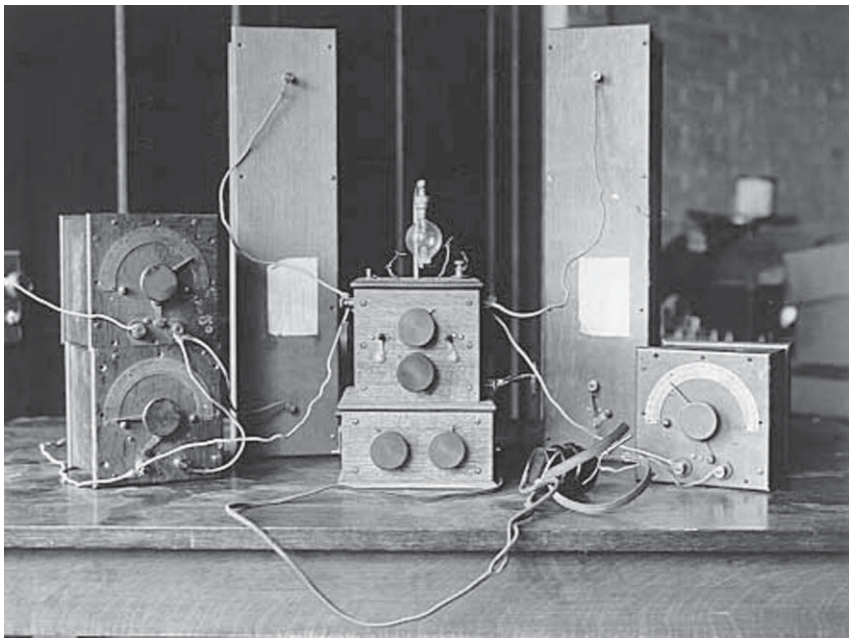
[FIG1] Junior telegraph operator David Sarnoff, Marconi Wireless Telegraph Co. of America, circa 1907. (Image courtesy of the Hagley Museum and Library [DS_1906].)

In early 1906, Sarnoff decided to try to get a full-time job (when he was about 15). Having sold newspapers, he decided that being a reporter, ultimately rising to editor or publisher, was his ambition, so he went off to the *New York Herald* to apply. However, he made an error. He asked a man behind a window in the lobby for a job, but that man was with the telegraph office of the Commercial Cable Company. The man advised him of his error and then told him of their need for a messenger boy and David took the job. It is hard to believe that the career of one of the most influential men of the early wireless industry was determined by such karma! Determined not to remain a messenger boy for long, he saw that the telegraphers were an important company asset, so he took two of his precious dollars and purchased a practice telegraph key. He was allowed to learn more by asking lots of questions and by working with some of the company telegraphers during off times. Nevertheless, this job did not last long. Sarnoff also sang in a synagogue choir for the high holidays to make a little extra money in the fall and asked for the days off. He was told to take them—and all the days following as well. He was fired!

Having become quite proficient with a telegraph key, he applied for a job as a junior operator at the seven-year old Marconi Wireless Telegraph Company of America—the American subsidiary of the British Marconi Company that controlled a great portion of wireless telegraphic messaging (Figure 1). He was hired as an office boy in September 1906, beginning his lifelong career with the single commercial entity that was to become RCA.

THE FATES OF YOUTH

In his junior year, Armstrong came under the influence of one of the founders of the Department of Electrical Engineering at Columbia, Michael Pupin. A Serbian immigrant, Pupin was the head of the Marcellus Hartley Research Laboratory and did his research in the basement of Philosophy Hall. His lab was full of the typical electrical engineering (EE) clutter



[FIG2] **The invention that started it all for Armstrong, the 1912 feedback (regenerative) receiver. This equipment was donated to the Smithsonian Institution following Armstrong's death. (Image from the Houck Collection site, reprinted with Permission from Mike Katzdorn.)**

and drew aspiring electrical engineers like a mecca. Pupin believed that engineers needed to be well grounded in basic science, quite at odds with most EE programs of his time. Pupin was also an inventor, having sold his patent for the Pupin loading coil, important for telephony, to AT&T for a goodly sum. In 1912, at age 21, Armstrong had an idea that would begin his inventing career. He could not wait to return to Philosophy Hall in September to develop it. On 22 September, he started a group of experiments that would lead to the regenerative circuit using the principle of partial feedback in a three-element tube (triode) to give amplification orders-of-magnitude more than what had been seen before (Figure 2). Moreover, when the amount of feedback was made too large, the circuit squealed (oscillated), which turned out to be an even larger discovery, for now there could be an electronic means for generating high-frequency waveforms, a necessity for both transmitters and receivers of wireless.

With his senior year still to finish and as he did more experimental work, he was sure that he had discovered the answer for which so many had been searching. While still pretty naïve about intellectual property and protection of it, he was careful enough to put his apparatus into a "black box" and not reveal what was inside. In the winter, he showed his remarkably clear reception of distant wireless signals to his radio friends from Yonkers but would not reveal the insides. He hinted at his accomplishments to his instructors who advised him to seek a patent. However, a patent cost US$150 to file and his only source, his father, refused to give him the money until he graduated in the spring. Father did not know best here! He sought out support from friends and relatives and even sold his motorcycle, but he could not come up with the full amount. However, an uncle who had some legal knowledge advised him to draw up a sketch of his device and get it dated and notarized, which he did on 31 January 1913. This cost him US$0.25.

AS A RESULT OF ALL THE NOTORIETY, THE COMPANY RECOGNIZED THAT SARNOFF WAS MORE VALUABLE THAN JUST A TELEGRAPHER AND STATION MANAGER.

Armstrong became absorbed in his experiments, taking some time to set up a second audion tube to show off its more controlled oscillating abilities, ultimately so important for transmitters and receivers too. While excitedly doing all this, he managed to graduate in June 1913 with a degree in electrical engineering; he was then offered an appointment as a teaching assistant in the department for one year at a salary of US$600—which he accepted. His father, prouder of the new job than his son's invention or graduation, came through with the money for the patent, which was ultimately filed on 29 October 1913. In this, his first experience with patents, he made an error that proved very costly later on. He insisted that any coverage for the oscillator function be sought in a second patent, not as part of the first one, as he wanted to develop the idea a little more in the laboratory. The oscillator patent was filed 18 December 1913.

In early 1907, 17-year-old David Sarnoff had all the duties of an office boy—dusting, cleaning typewriters, and emptying wastebaskets—but he managed to sometimes practice his telegraphy communicating with the four Marconi stations and with Western Union. Soon he was trusted to receive messages. His superiors found it amusing that he read many of the correspondences he was to file—he was learning about the business and at the same time improving his English. He also spent time with the laboratory technician and, after blowing out dozens of fuses, was able to set up experiments and do repairs in the lab. An exciting period was when Marconi himself visited New York and spent some time in his Front Street workshop. David lugged his suitcases, delivered candy and flowers for him, and experienced the aura of the great inventor. Apparently, Marconi saw in David a first-rate mind; the man and the boy engaged in very philosophic discussions. A few months after turning 16, Sarnoff was promoted to "pony operator" (junior telegrapher) at the salary of US$7.50 per week. A move to Nantucket Island to be one of the four operators of that station

earned him his next promotion at age 17 to "assistant operator" at US$60/month and soon thereafter to "full operator" at US$70/month. On Nantucket, he acquired a second-hand bicycle, often pedaled the seven miles to the Nantucket Library, and took correspondence courses in algebra and geometry.

In 1909, Sarnoff was made manager of Marconi's Coney Island station, the company's busiest. While taking a US$10 a month cut in salary (he was back in New York after all) he was becoming well-known worldwide by telegraph operators who could recognize his "fist" (the rhythmic patterns unique to a telegraph operator). In 1911, Sarnoff took an assignment on the sealing ship, *Beothic,* and was involved in the first time radio was used to aid in a medical emergency. He also served as the telegraph operator on the SS *Harvard.* He then was given the task of manager and operator for the 5 kW (most powerful station in New York) station atop Wanamaker's Department store. With a second station located in their store in Philadelphia, Wanamaker's hoped the novelty of radio would attract thousands to the stores. For Sarnoff, finally having regular hours allowed him to take a special night course in electrical engineering at Pratt Institute that compressed three years into one.

Even his diligence and skill, however, could not push his career as well as another event (which was fortuitous for him). The 21-year-old Sarnoff was at the key at Wanamaker's on 14 April 1912 and received weak and static-filled radio messages from the SS *Olympic*, saying that the *Titanic* had hit an iceberg and was sinking fast. For the next 72 hours Sarnoff stayed at his post, acknowledging the messages and relating the details to the public and the assembling crowd that converged on the store. President Taft ordered all other U.S. stations to power down. Not until the complete list of survivors was given to the press did he quit. Sarnoff had been at the key position and had taken proper action, showing all too clearly that many more lives could have been saved had there been more and better radio apparatus aboard vessels. Congress immediately passed the Radio Act making it mandatory for all ships carrying more than 50 persons to have and continuously man a radio system. The Marconi Wireless Telegraph Company of America began to make some money!

As a result of all the notoriety, the company recognized that Sarnoff was more valuable than just a telegrapher and station manager. Near the end of 1912, he was appointed its radio inspector for ships in New York Harbor and, a few months later, the chief inspector for the whole country. He was also given the title of assistant traffic manager, and he served as an instructor at the Marconi Institute, a training school for radio operators. By the summer of 1914, Sarnoff was named the contract manager for the new American director of the Marconi Wireless Telegraph Company of America, Edward J. Nally. There is little doubt that his influence was growing well beyond whatever title he now held.

THE FIRST MEETING

In December 1913, a demonstration of Armstrong's new apparatus was set up in Prof. Pupin's laboratory for a group from the Marconi Wireless Telegraph Company of America. By this time, David Sarnoff had the position of assistant chief engineer, and on 6 January 1914, he came with two older Marconi engineers. The two 23-year-old principals, each of genius talent, appeared and acted very differently: "Armstrong tall, slow-spoken, extremely reserved, with an analytical mind; Sarnoff short, fluent, aggressive, the entrepreneur" [7, p. 53]. Surprisingly, they hit it off. Sarnoff could see the potential for the commercialization of what Armstrong showed, likely more so than did his older colleagues. The Columbia "wizard," who still kept his apparatus in a covered box, showed clear reception of signals from Marconi's transmitter in Clifden, Ireland. The transcriptions were verified a few days later by the company. Sarnoff's report the next day might be considered a turning-point document, saying, while he would like to see further tests, "the results obtained were, I thought, quite phenomenal" [9, p. 65].

ALAS, THE CUT-THROAT WORLD BEGAN TO SHOW ITSELF TO ARMSTRONG JUST AFTER HE HAD FILED FOR HIS PATENT IN LATE 1913.

About two weeks later, Armstrong was asked to demonstrate his apparatus at a new Marconi transmission station in Belmar, New Jersey. Armstrong and Sarnoff went through the night in a drafty wireless shack on a bitter and cold night copying messages from Clifden, Cornwall (England), Nauen (Germany), and Honolulu, which all came in strongly in the early morning hours. Again, all these were verified. Sarnoff's letter based on this experience was stronger, "the most remarkable receiving system in existence" [9, p. 65]. He recommended acquiring the system without delay. When the word of this recommendation reached Sir Godfrey Isaacs, the managing director of the Marconi Company, his temper flared, saying to the effect, that this young upstart should be fired, being so ready to spend the company's money! The Marconi Company did not buy into Armstrong's system at that time and Sarnoff was not fired. What is clear is that the two men always looked back fondly on that night in which so much about the possibilities for wireless became evident.

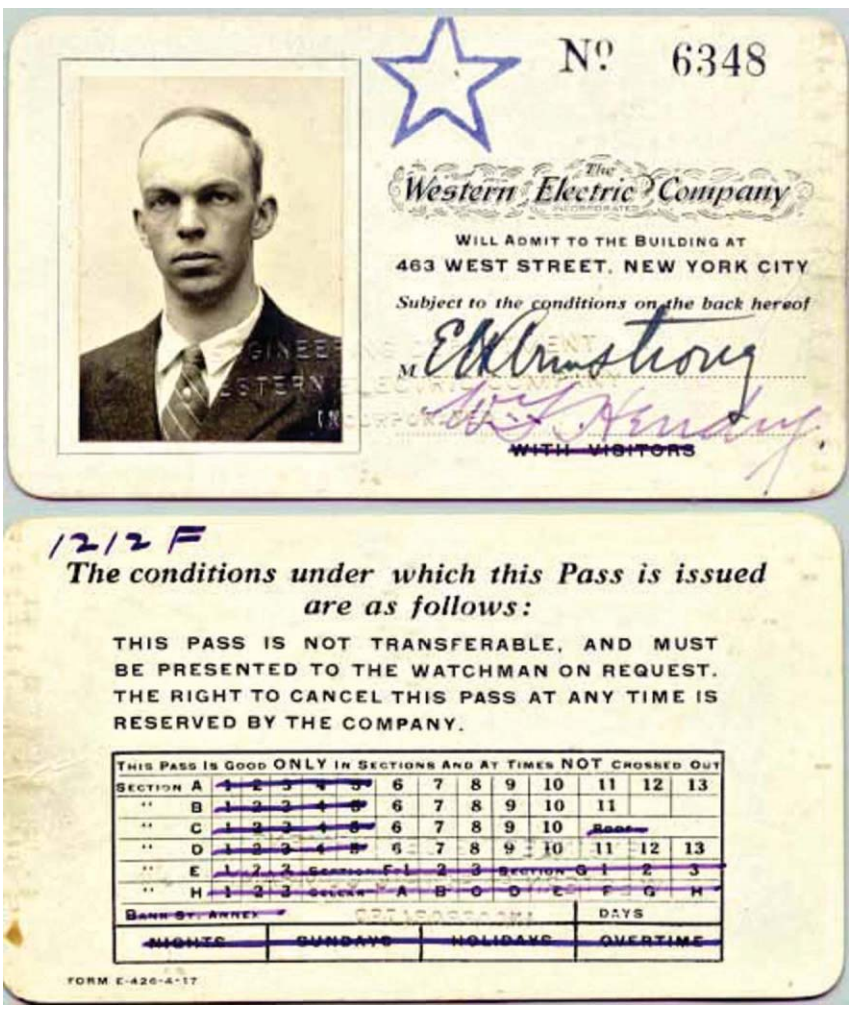
PREWAR

In February 1914, Prof. Pupin bragged to a group at the University Club in New York City that he had listened to Honolulu. The chief engineer of AT&T, J.J. Carty was there and very skeptical, so a demonstration was arranged for some of their technical personnel. One should note that AT&T by this time had acquired the rights to the de Forest audion for wire and telephone use via some rather devious [6, pp. 108–109] indirect purchase by an attorney and some desperation on the part of de Forest, who had just had one of his companies go bankrupt and was on the verge of being indicted for stock fraud. AT&T was working extensively on the audion and Fritz Lowenstein, an AT&T inventor, had just used the audion to amplify in a telephone circuit without

screeching. How dare this unknown college student do better! After disclosing the workings of his circuit in the spring of 1914, Armstrong did not hear from that company again.

Somewhat more illuminating is Armstrong's first meeting with Lee de Forest at the 1913 meeting of the one-year old Institute of Radio Engineers (IRE, one of the predecessors of IEEE). De Forest lectured at Columbia about his audion and its use in a telephone repeater circuit and really wanted to see if the rumors he had heard about a Columbia student's discoveries were true. Armstrong, however, knew that de Forest had been admonished a few years before for trying to patent, essentially, a detector that he had seen in Fessenden's laboratory, so he was willing only to give a "black box" demonstration. There was little doubt that the two men disliked each other at first sight.

In the summer of 1914, Armstrong's instructorship ran out, but a one-year fellowship was found and he continued work at Columbia with Prof. Pupin attacking, for the first time, the problem of static in wireless systems. On 6 October 1914, his first patent was issued on the regenerative receiver circuit. With the outbreak of the war in Europe, the British cut off all undersea cables to Germany, leaving Germany only with wireless communications to the United States. After a demonstration of his apparatus in October 1914, the German company Telefunken and the German Embassy used Armstrong's receiver and was the first to license his patent, paying fees until the United States entered the war in 1917.



[FIG3] **(a) Front and (b) back of a Western Electric Engineering Department building pass issued to Armstrong before World War I. With his growing reputation among the radio community, and his many professional contacts at the IRE and The Radio Club of America, he must have been a frequent visitor to labs such as this one. (Image from the Houck Collection site, reprinted with permission from Mike Katzdorn.)**

In late 1914–early 1915, Armstrong, with the help of Prof. Morecraft at Columbia, published and gave talks about the properties of the three-element tube (triode or audion) that clarified its use as an amplifier and/or oscillator. These were the tools that electrical engineers needed to design these devices into various products. De Forest, at about this time was ridiculing Armstrong's smoothly drawn curves. He strongly argued that the curves needed to have some "wiggles" in them. The discussions, which were reported in the IRE journals, were quite heated; history has shown that it was clear that de Forest even these years later still did not understand how his invention worked!

By 1914, Sarnoff had become the "de facto if not the de jure most effective adviser on technical and commercial policies" [9, p. 68] for the company. He was becoming its spokesman as well. His ideas for tolls, rates, and routing raised some controversy when he presented a paper to the IRE, but ultimately it was his views that prevailed. His opinion as an "out of the box" prognosticator was presented to Vice-President Nally in 1915 when he wrote him a memorandum developing a case for his in infamous "radio music box." Sarnoff realized that one of the problem issues with his company's business was that the radio transmissions were open to the public and not private for the sender and receiver. He thought this could be turned into an asset by using "broadcasting" to send out to thousands, if not millions, of receivers owned by the public. This was five years prior to the first true broadcasting station in the United States. In the memo, Sarnoff reasoned that they could bring the cost of the receiver down to about US$75 and sell a million of them. While the numbers were small compared to what really happened, the idea was proved by history. Marconi management read the memo in "wide-eyed amazement,"

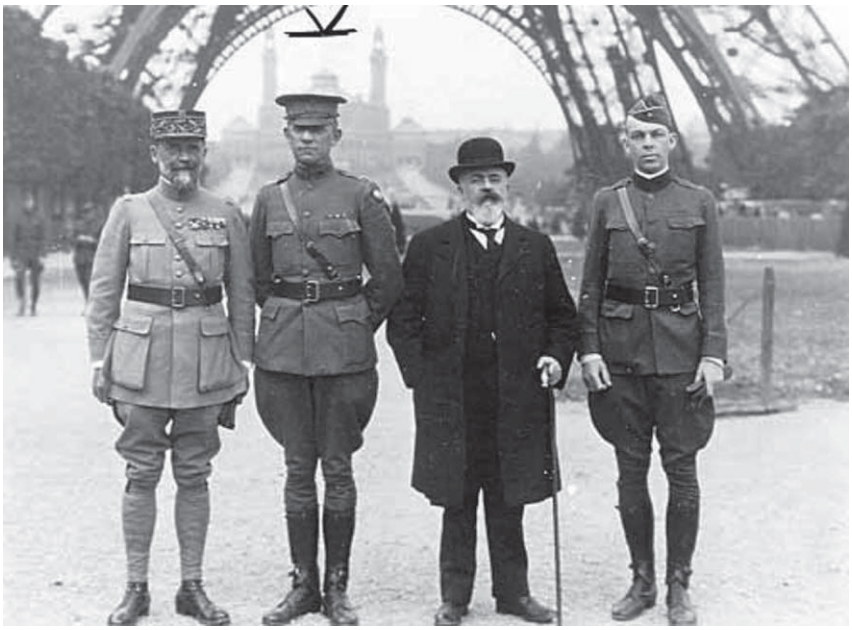
[9, p. 73] considered it harebrained, and it was filed and forgotten.

By 1 January 1917, Sarnoff had been promoted to the commercial manager of the growing company, one that was tripling its manufacturing capabilities. He was widely recognized as a spokesperson for the wireless industry and was made the secretary of the IRE. He had learned to interpret very intricate issues into a few basic tenets and had dealt with patent entanglements, licensing, and commercial rates. While not an inventor himself, he was very current in the technical intricacies of the industry and "his judgments on the enfolding science were not warped by the kind of ego drives and emotional reactions that unavoidably affected the views of men like de Forest, Fessenden, and Armstrong" [9, pp. 74–75]. Also, in February 1917, Sarnoff married Lizette Hermant, a vivacious French-born woman.

Alas, the cut-throat world began to show itself to Armstrong just after he had filed for his patent in late 1913. Three claims were made by others for the same or similar invention. One was from GE's excellent scientist Irving Langmuir, who was known for great improvements to the vacuum tube, but who had arrived at regeneration independently some time after Armstrong. Next was by a German, Alexander Meissner, who filed on 16 March 1914 for a regenerative circuit for a gas discharge tube. The third was Lee de Forest, who filed two patents, the first of which was dated March 1914, about six months after his "interaction" with Armstrong. All these were used for interference proceedings with Armstrong.

WORLD WAR I

In 1917, Armstrong, a Theodore Roosevelt progressive, Republican, and Protestant Presbyterian, joined into the patriotism of the era. He was quite famous among the radio community and had just been elected the president of the Radio Club of America in 1916 (Figure 3). He decided to join the Army and was commissioned a captain in the U.S. Army Signal Corps, given short training in the summer of 1917 and had the good fortune to be stationed in Paris for the duration. At first he helped build some equipment that was devoid of vacuum tubes and his regenerative circuits. Then he was assigned to work on radio systems for communicating with aircraft. Of course, this gave him ample excuses to fly along with the equipment, which pleased him immensely.



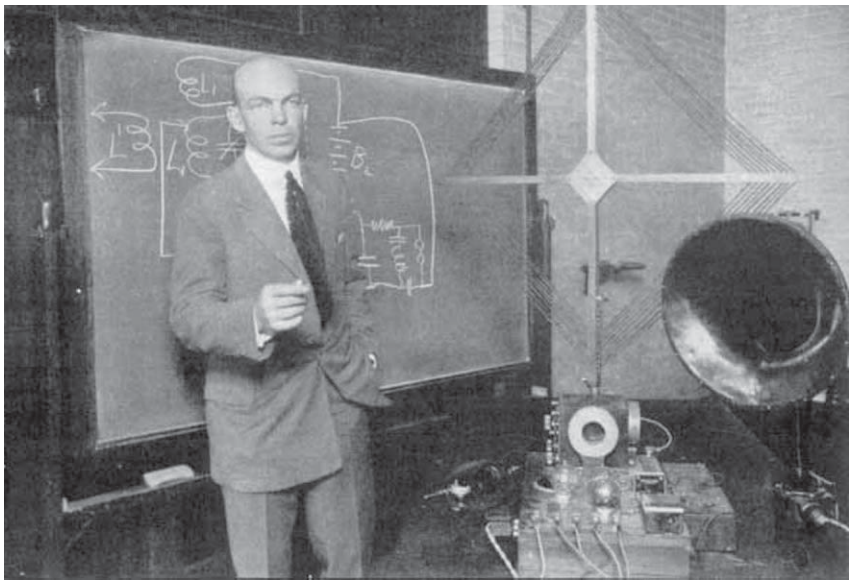
[FIG4] From left: General Ferrie, an unidentified man thought to be his aide, Prof. Abraham, and Armstrong in front of the Eiffel Tower, 1918. (Image from the Houck Collection site, reprinted with permission from Mike Katzdorn.)

One evening in Paris, at a time when the German air force stepped up its bombing of the city, Armstrong was out watching the "fireworks" from a bridge over the Seine. He wondered whether he could somehow detect the very high-frequency (for that time) ignition noise electromagnetic emissions from the airplanes and use them in a direction finder to guide anti-aircraft fire. He had been dealing with the issue of how to detect these high-frequency waves with the currently available vacuum tubes that could not yet amplify well at these frequencies. At one intersection on his way back to his apartment, the idea hit him for what was to become the superheterodyne system. He immediately hacked together an eight-tube receiver. "Where the regenerative circuit amplified signals up to about one thousand times, the new circuit increased this by several thousandfold, with unprecedented stability, selectivity, and quietness" [7, p. 70]. He was to apply for a patent on this, but the application was dated 30 December 1918, signed in Paris. It reached the U.S. Patent Office on 8 February 1919 and was issued very quickly on 8 June 1920. This may have been his most important contribution, but it was not the patent that caused him to have his many patent legal battles.

In February 1919, he rose to the permanent rank of Major for his distinguished service. He was asked to give two lectures at the Sorbonne: one on regeneration and the other on the superheterodyne, and was awarded the Chevalier de la Legion d'honneur at the Palace of Justice (Figure 4).

He was unable to get home until late in September 1919 and in absentia was awarded his most prized honor. The IRE (the larger of the two societies that became the IEEE in 1963) had awarded its first Medal of Honor to him for his invention of the feedback circuit. This award was to play a role in Armstrong's later situations.

Sarnoff, having had many interactions with the Navy, the premier service for military radio, also immediately went down to the Brooklyn Navy Yard, applying for a commission in their



[FIG5] **Edwin H. Armstrong explaining the principles of his latest invention, "superregeneration" at a meeting of the RCA held at Columbia University, New York City. (Image from the Houck Collection site, reprinted with permission from Mike Katzdorn.)**

communications area. However, his commission was blocked by "race prejudice in Washington" [9, p. 77]. He refused to ask his local draft board for a deferment and was thus certified for active service in the Army. However, Admiral R.S. Griffin wrote to the board that Sarnoff was essential in his current position, "in that the Fleet will not suffer delays due to unsatisfactory deliveries in existing contracts" [9, p. 77]. So Sarnoff performed his wartime duty in New York. The Germans and the British had succeeded in cutting each other's transocean cables, so wireless was more vital than it had been prior to wartime. In fact, President Wilson signed an order that all radio facilities, both commercial and amateur, be taken over by the Navy in April of 1917. Nevertheless, under Sarnoff's principal guidance, the Marconi Company had unprecedented sales of over US$5 million in 1917. He had made contacts with virtually all those in the military who dealt with communications (which would later prove important) and received many commendations for his leadership during this period.

THE POSTWAR PERIOD

In 1919, the British Marconi Company wanted to acquire minimally, unrestricted access to the Alexanderson high-frequency alternator, the rotating machine, high-power transmitter that was developed and controlled by GE. As early as 1916, Sarnoff had written a position paper for the Marconi Company that the entity that had this technology would have a substantive advantage in the wireless communications business. Three months after the armistice, British representatives came to complete the multimillion dollar deal. The Navy, however, had alerted President Wilson that having foreign control over wireless communications in the United States was a bad idea. Assistant Secretary of the Navy, Franklin Roosevelt, then wrote to GE to postpone the sale of the transmitters and a conference was held in New York on 8 April 1919. The head of GE's legal department, Owen D. Young, was then asked to direct a group that was to let GE forcibly "buy out" American Marconi. GE completed the sale of the American Marconi interests on 20 November 1919 for over US$3 million and formed the RCA.

However, there was going to be a problem when the government was to allow wireless to return to private industry; the patent structure was very messy and no one company could build anything without infringing on patents held by another. AT&T held the de Forest patents and had accumulated many important patents, (principally on improvements for vacuum tubes) over the previous ten years, so a few months after RCA was formed, AT&T became one of the principal owners, purchasing about US$2.5 million in RCA stock. AT&T was to concern itself with all radiotelephony associated with telephone service and the manufacturing of transmitter equipment, while GE was to control all wireless telegraphy and receiver equipment construction.

The other large electric corporation, Westinghouse, did not want to be left out of the new arrangement. It had a modest patent portfolio of its own, but soon added to it by purchasing the International Radio Telegraph Company, the nearly bankrupt successor to Reginald Fessenden's National Electrical Signaling Company and later on purchasing the rights to all of Armstrong's patents. When Armstrong returned to the United States in the fall of 1919, his only income from licensing was the US$5,700 that had been paid by the Marconi Company over the previous two years. Thus, while always reticent to sign over patent rights, Armstrong was convinced by his patent attorney to sign over both the feedback and the superheterodyne rights to Westinghouse for US$335,000 payable over ten years, and for an additional US$200,000 payable when his oscillator patent was cleared from de Forest's pending interference suit. With these in hand, Owen Young was able to add Westinghouse to the ownership of RCA effective mid-1921. The deal was that Westinghouse was to do 40% of the manufacturing for RCA while GE was to retain 60% for itself. RCA was to have no manufacturing rights.

Owen D. Young became the first chief executive officer of RCA, Edward G. Nally its first president (he had been the managing director of American Marconi), and GE's Alexanderson the chief engineer. Sarnoff retained his position as commercial manager but was not named to the board of directors. In the beginning, only the memo justifying the idea

of a "radio music box" that Sarnoff had submitted to the Marconi Company had anything to do with RCA being in broadcast radio. However, in early 1920, Dr. Frank Conrad, a researcher at Westinghouse, started some regular broadcasts for amateur (typically crystal-set builders using earphones) listeners from his home in Pittsburgh. He was able to convince his company to build a new transmitter at Westinghouse and to broadcast the election results of 1920. Thus was founded what is accepted as the first radio station, KDKA in Pittsburgh, which received its license on 27 October 1920. It was to broadcast on a clear channel of 360 m (833 kHz). What followed was a stampede like no other before it. By 1 May 1922, about 18 months later, 218 [4, Exhibit G] of what we would today call commercial broadcasting licenses had been granted. However, with no reasonable regulation, most of these were crowded into a few frequencies. Given interferences and receivers that tuned badly, amateurs, and commercial messaging, it was a mess!

Armstrong, while devoting a lot of nervous energy to several litigations regarding all his patents, went back to Prof. Pupin at Columbia, working closely with him trying to solve the problem of radio interference for the standard amplitude modulation (AM) transmissions. He would work on this problem for the next 12 years before the wideband frequency modulation (FM) solution was found, and in the course of the research had many false starts based on inaccurate hypotheses. The wartime experience had impacted Armstrong, and he had returned as a more confident 28-year-old and was established as an expert among the growing group of radio amateurs. In 1920, de Forest had reopened his station in New York, broadcasting gramophone records and even using some live talent, but the Navy closed him down, since he was interfering with their messages. Dozens of amateur stations were also available, and electronics parts shops exploded with radio essentials, mostly for crystal sets. Armstrong had kept the rights to license his patents to amateurs and many wanted to use his regenerative circuits, so that by 1922, he was earning about US$10,000 per month from these licenses alone.

With the terrible crowding in the "normal" radio band, in 1921 the government restricted amateur radio transmissions to the "commercially useless" short wave band of over 1.5 MHz! In the summer of that year *QST*, the magazine of the American Radio Relay League proposed a test for this frequency range, wanting to obtain highly reliable transmissions with Great Britain. Armstrong and his Radio Club, using a transmitter of his own design and a superheterodyne receiver setup in Greenwich, Connecticut, not only made strong contact with the group specifically set up to receive the "high-frequency" commercial transmissions, but also with amateurs from Holland, Germany, and Catalina Island in California. This group of amateurs had done for about US$1,000 what industry might have done with greater ease, should it have been so inclined. Of course, Sarnoff was not one to miss the event. He and a group of RCA aides all came out to see what had been considered "impossible." Still, it took until about 1927 before Congress legislated to use some of these "short-wave" frequencies to get some semblance of order for the crowded airwaves. Stations continued to try to drown out their competitors and, if so inclined, worked out arrangements so that only one was on at a time.

It was also in 1921 when Armstrong, preparing to testify at the interference hearing against de Forest, was going over some of his old regenerative circuitry to refute a fact brought into evidence by an opposing lawyer. Suddenly, he heard reception so clear that it was very much beyond what he would have expected from the regenerative circuit. He immediately concluded that he no longer fully understood all about regeneration and subsequently invented a new system that used a second tube to "quench" the annoying squeal (oscillations) about 20,000 times a second. This allowed an amplification of up to 100,000 times the original signal strength, even more than for the superheterodyne. Now, being experienced with patents, he filed for a patent on superregeneration in 1921 (see Figure 5) and received a patent in July 1922, with six more following on the idea soon thereafter. Armstrong's timing could not have been more perfect, as this was at the beginning of the broadcasting explosion. Sarnoff was among the first to see the new system.

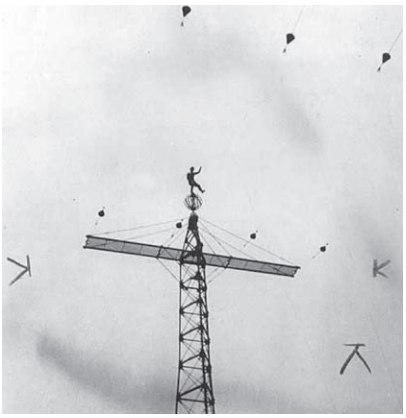
As RCA started to function, Sarnoff had the challenge of working through the maze of licenses and agreements with the large corporate shareholders of RCA all the while developing the wireless communications business and the burgeoning of the radio broadcast industry. As no corporate entity had foreseen broadcasting's spectacular growth, RCA and its corporate owners had to reach for a part of this new business, often in competition with one another. Keeping peace among them was no easy task, and Sarnoff worked endless hours on agreements, litigations, marketing, sales, and planning, having gained great friendship and trust from Owen D. Young.

From their first historic meeting in that Columbia laboratory in 1914, Sarnoff and Armstrong remained friendly. After the war, it was not unusual for Armstrong to drop in to Sarnoff's apartment at breakfast time for "just a cup of coffee" [9, p. 112]. Sarnoff's children called him "the coffee man." They were also often together at the RCA office discussing Armstrong's research and plans. While visiting in 1922, Sarnoff's secretary, Esther Marion McGinnis, a tall "strikingly handsome" [9, p. 112] woman of 22, who was out on her own in New York, abandoning the factory town of Merrimac, Massachusetts, caught Armstrong's eye. Lively and witty, she politely deflected Armstrong early advances. However, after Armstrong took a trip to France in October 1922, bringing back a fancy Hispano-Suiza automobile, she accepted a ride the next spring. The affair was on.

About this time, RCA was constructing its first New York broadcasting station, building twin 100-ft towers on top of the Aeolian Building on 42nd Street. There were crossarms on each tower on which a man could walk and, 15 ft higher, a large iron ball. Armstrong, always a bit of a daredevil, enjoyed being in high

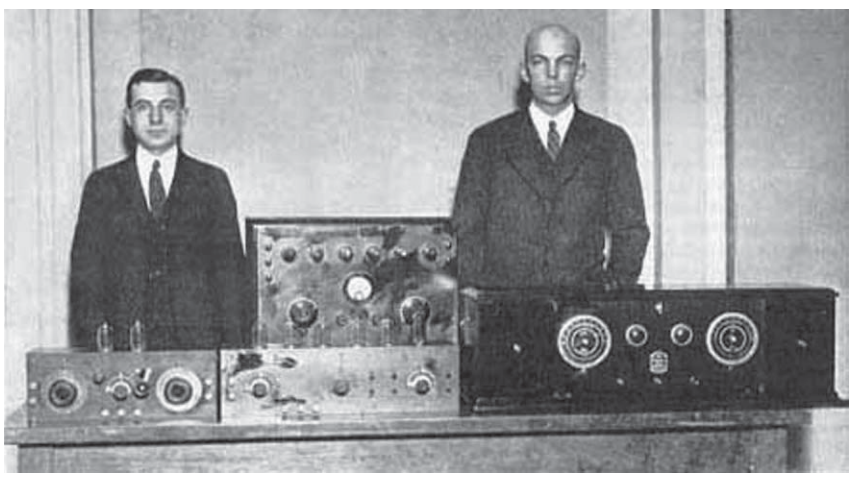
places, so he often came to watch the construction; however he also enjoyed climbing up and dangerously balancing himself atop the iron ball. This came to Sarnoff's attention, and he wrote Armstrong a strong letter to stop doing this. However, on the station's opening day, Armstrong climbed the tower and posed for the well-known picture shown in Figure 6, some 350 ft above the street. While this led to Sarnoff banning Armstrong from the station, it did not affect their friendship. Shortly thereafter, Marion McGinnis accepted Armstrong's proposal for marriage. After a harrowing trip in his new car from New York to the Boston area in December 1923, they were married in Merrimac and tried to take the Hispano-Suiza for their honeymoon to Florida. However, the car (and they) traveled more by train than they had expected! They moved into an apartment on 86th Street and Riverside Drive.

The friendship, inventions, and the need all came together in 1922 and 1923. RCA, unable to manufacture on its own, had a very slow development process that killed sales of RCA radio sets as other companies could offer the latest developments. While crystal sets were at first the most widespread, new designs using vacuum tubes were appearing. One in particular, the Neutrodyne, invented by Alan Hazeltine, a friend of Armstrong and a professor at Stevens Institute of Technology in New Jersey, was widely



[FIG6] **Armstrong atop the new WJZ/ WJY Broadcasting Tower in New York City, 15 May 1923. (Image from the Houck Collection site, reprinted with permission from Mike Katzdorn.)**

accepted by the public. It successfully avoided the large patent entanglements controlled by RCA, and its design was put up by Hazeltine for licensing to a large number of RCA's competitors. RCA's offerings, often six months behind, had a very poor share of the marketplace. Thus Sarnoff saw an opportunity for RCA, having been so impressed with the abilities of the superregenerative system. Of course, first he needed to do "due diligence" as the corporate manager, so he had his patent department search for a patent with which RCA could challenge Armstrong's invention. They only found one marginally similar English patent granted to a man named John Bolitho. However, Armstrong had done his homework and had purchased the rights to the patent for a few thousand dollars. Ultimately, Sarnoff's attorneys were told if they wanted the Bolitho patent to see a man named Armstrong. Thus, Sarnoff persuaded his board to offer Armstrong US$200,000 and 60,000 shares of RCA stock for the exclusive rights to superregeneration. This was a very large sum for the time. RCA stock was booming, and this transaction made Armstrong an instant millionaire. Unfortunately for Sarnoff, in a short time RCA and Armstrong concluded that the superregenerative system could not be mass produced easily. Superregeneration was used for identification, friend, or foe (IFF) transponders, police radio, ship-to-shore, and other special purposes later on. Despite its magnificent amplification, it could not be made sufficiently frequency selective, dooming it for the radio industry of that time. However, Sarnoff seemed to always have a lucky streak; for, at the same time, he had also obtained exclusive rights to Armstrong's superheterodyne system. Armstrong had demonstrated his latest design at then Vice-President Sarnoff's apartment in early 1923. Sarnoff was very impressed, having been told by his own technical staff that the superheterodyne design was years away. Armstrong also took another set to the apartment of RCA Chairman Owen D. Young, entering with the battery-powered receiver blasting an opera that had clarity beyond anything he had ever heard.



[FIG7] **Houck and Armstrong with three superhets: (a) The Signal Corps. set built in France, (b) the preproduction second harmonic superhet, and (c) the production second harmonic superhet, the Radiola Superheterodyne, 1923. (Image from the Houck Collection site, reprinted with permission from Mike Katzdorn.)**

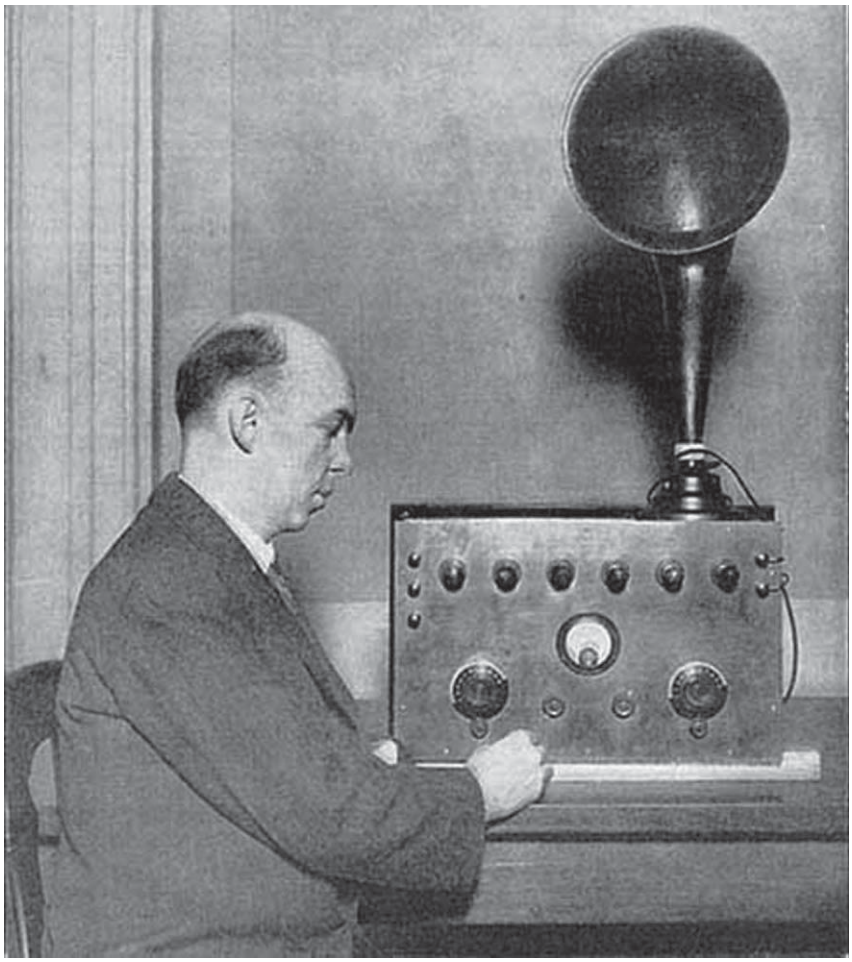
At the time that the licensing arrangement with Armstrong was consummated, RCA was about to conclude its yearly commitment for several million dollars with its two manufacturers for radio sets who were to build an improved design developed by the RCA technical staff. With the superheterodyne now in hand, Sarnoff canceled the next meetings needed to conclude the contract. He was willing to sacrifice a year of sales to restart with the technically superior product. The RCA staff began a crash program to get the superheterodyne system to market in 1924, hopefully leapfrogging its Neutrodyne (et al.) competitors. By midsummer of 1923, Sarnoff heard that AT&T (part of the licensed group) was

preparing to enter the radio receiver business with a superheterodyne receiver of its own and was going to install a set in the White House for President Harding. However, Sarnoff's staff was having a lot of problems developing the superheterodyne for production.

When Dr. Albert Goldsmith, RCA's head of their laboratory came to Sarnoff to advise that the project be abandoned and the original contracts reinstated, Sarnoff was in a tough spot. Seeing the crisis, Marion McGinnis suggested, "Why don't you call Armstrong?" [7, p. 119].

Not only had Armstrong made further developments of the superheterodyne, but, Harry Houck, working with him (see Figure 7), had developed technology sufficient for a patent for using the second harmonic in the superheterodyne receiver and had reduced the complexity of the receiver to two knobs. After a few intensive weeks, and successfully getting the new RCA receiver ready for production, Sarnoff awarded Armstrong an additional 20,000 shares of RCA stock for pulling off this success. This proved to be a good investment, as the new receiver, introduced for sale in early 1924, was a sensation and put RCA far ahead of its competitors (Figure 8). RCA dominated the market for the next three years, also working hard to try to mitigate the Hazeltine patents. Surprisingly, the additional shares made Armstrong the largest single stockholder in RCA. Finally, in 1927, a system of more liberal licensing was put into place that allowed virtually everyone to use the patents, making the "superhet" available to all manufacturers. RCA received modest royalties under the agreement that it had to share with the electric companies. Armstrong was never paid any royalties by RCA.

After the licensing agreement was concluded with Westinghouse in 1920, the company backed his decision to prepare a suit against the de Forest Radio and Telegraph Company, one of about 25 [10, pp. 86–87] companies that de Forest started that ultimately went bankrupt. This suit, to once and for all mitigate de Forest's infringing claims, started hearings in the federal court of the southern district of New York in January 1921. This



[FIG8] Armstrong and his prototype second harmonic superheterodyne (radio broadcast 24 July 1923). (Image from the Houck Collection site, reprinted with permission from Mike Katzdorn.)

fight ultimately took 14 years and had a severe impact on Armstrong. This is well expressed by Lessing, who quoted one of Armstrong's patent attorneys:

> ...there were really three Armstrongs, closely related, but distinct, contained in the same character but separated by the increasingly harsh attrition of events. The first of these was the easy, modest man of the private world who could relax swiftly among friends into intercourse of intangible charm and grace. This private man held a legion of friends over the years, with but one notable disaffectation [de Forest]. The second Armstrong was the man of finance and affairs, who could press forward aggressively to achieve his ends. A little apart from these figures stood Armstrong the inventor, proud and lonely and raised on a pedestal, whom the other Armstrongs protected with the jealousy of a brother and whose life went forward in the intensely individual processes of creation, oblivious for long periods of even wife and friends. At the slightest threat to the image of the inventor, all the forces of the whole personality marshaled to repel the attack.

In 1922, Armstrong's patents had been sustained twice, both by the District Court and the Second U.S. Circuit Court of appeals. It was clear from the testimony that the true inventions were Armstrong's, as the IRE had concluded in its first Medal of Honor Award in 1919. It was, however, at this point that these apparent legal victories started to unravel. In a patent case such as this, no

final judgment can be entered until either damages are drawn up or the winner has waived damages. The stubborn Armstrong refused to waive damages, against the advice of his supporting Westinghouse attorneys. Meanwhile, the special master assigned to assess damages died suddenly, there were great delays in appointing a successor, and the de Forest Company went bankrupt. In 1924, out of this maze of litigation, de Forest was able to get a decision from the District of Columbia Court of Appeals ordering the Patent Office to issue him immediately patents on both his two patent submissions (applied for later than Armstrong's and after he had seen Armstrong's apparatus) for the "ultraaudion" and his regenerative system. Quoting from a private conversation with my Brown University office neighbor, Barrett Hazeltine, the son of Armstrong's friend, Alan Hazeltine, "My father always said de Forest never understood his own circuit." Now the battle was on again in an interference suit that turned more to legalisms and word games than true technical evidence. Also, de Forest's interests were defended by all of AT&T's legal department, while Westinghouse, whose advice had been spurned by Armstrong, no longer supplied its legal force. Two men now held the same patents. Several suits were filed in Delaware. De Forest sued to have Armstrong's patent declared null and void in Philadelphia. With the issue thus confused, AT&T and de Forest won in both places and in 1927 these findings were upheld by the Third Circuit Court of Appeals.

The losses to de Forest were taken as a personal affront by Armstrong, so he decided to carry the battle to the Supreme Court. This he did in 1931. Marion Armstrong worked hard for her husband, as did his friends from the Radio Club. This was the topic for many discussions at IRE meetings. After some wins and losses in lower courts the final verdict by the Supreme Court was rendered on 24 May 1934, siding with AT&T and de Forest. By this time, RCA had it in its own interest to back the de Forest suit as well, which was a first crack in the friendship between Armstrong and Sarnoff. While there were petitions from Michael Pupin, Alan Hazeltine, and other of Armstrong's supporters who believed the verdict was a total distortion of the scientific facts, the petition was denied. On 28 March 1934, the IRE held its ninth annual convention at the Hotel Benjamin Franklin in Philadelphia. Armstrong walked into the meeting of some 1,000 engineers who were stunned to see the inventor that had been so poorly treated by the patent system. The Institute had been informed that he would like to give a speech, returning the 1918 Medal of Honor awarded to him for the discovery of regeneration. The speech in his pocket began [7, p. 153]:

> It is a long time since I have attended a gathering of the scientific and engineering world—a world in which I am at home—one in which men deal with realities and where truth is, in fact, the goal. For the past ten years I have been an exile from this world and an explorer in another—a world where men substitute words for realities and then talk about the words. Truth in that world seems merely to be an avowed object….

The speech was never presented. IRE President Charles M. Jansky went to the podium and announced that, by a unanimous decision of the Institute's Board, the medal had been correctly awarded. He reaffirmed the citation for his discoveries of regeneration and oscillations for vacuum tubes. He was given a standing ovation. This is especially poignant in that half of the Board members were employed by AT&T, RCA, or their affiliated companies, and this was the depth of the depression. Nevertheless, the corporate departments responsible for putting out press releases now worked hard to build up de Forest as "the father of radio." While somewhat buoyed by the decision of the IRE, the patent fight took its toll on Armstrong continually, from that point on, narrowing his group of trusted individuals.

> For a viewpoint totally out of agreement to what is presented here, one may read the letter written by Lee de Forest in 1956 as a criticism to a positive article on Armstrong published in *Harper's Magazine* in April 1956, which is preserved with author Carl Dreher's response near the end of the Houck Collection.

In the second part of this article, we shall finish the story, observing that, as the goals of the two giants diverge, so does their relationship. Armstrong, while hurt by losing his long patent battle for the regenerative oscillator, still has another breakthrough to reveal in the 1930s—FM radio. Sarnoff, while having to deal with fiscal problems of the depression of the 1930s, has one of his largest successes, using RCA as the driving force for television and later even wins for his standard for color television. However, all the successes did not necessarily lead to happy endings.

I hope you are looking forward to reading the story's conclusion in the January issue of *IEEE Signal Processing Magazine*.

AUTHOR

Harvey F. Silverman (hfs@lems.brown.edu) is a professor of engineering at Brown University.

REFERENCES

[1] G. L. Frost, *Early FM Radio*. Baltimore, MD: Johns Hopkins Univ. Press, 2010.

[2] S. Fybush. (2005, June 10). *Tower Site of the Week* [Online]. Available: http://www.fybush.com/sites/2005/site-050610.html

[3] L. Gleason and L. Archer, *Big Business and Radio*. New York: American Book–Stratford Press, 1939.

[4] L. Gleason and L. D. Archer, *History of Radio to 1926*. New York: American Book–Stratford Press, 1938.

[5] D. G. Godfrey, *Philo T. Farnsworth the Father of Television*. Salt Lake City, UT: Univ. Utah Press, 2001.

[6] M. Katzdorn. (2010, Sept. 7). E. H. Armstrong Web Site. [Online]. Available: http://users.erols.com/oldradio/index.htm#Y

[7] L. Lessing, *Man of High Fidelity: Edwin Howard Armstrong*. Philadelphia, PA: Bantam Books, 1969.

[8] H. M. Library, "David Sarnoff circa 1907," *The David Sarnoff Library Photograph Collection*, Accession 2010.271, 2010.

[9] E. Lyons, *David Sarnoff*. New York: Harper & Row, 1966.

[10] W. R. Maclaurin, *Invention and Innovation in the Radio Industry*. New York: Macmillan, 1941

[SP]

Zhou Wang

applications CORNER

Applications of Objective Image Quality Assessment Methods

The interest in objective image quality assessment (IQA) has been growing at an accelerated pace over the past decade. The latest progress on developing automatic IQA methods that can predict subjective quality of visual signals is exhilarating. For example, a handful of objective IQA measures have been shown to significantly and consistently outperform the widely adopted mean squared error (MSE) and peak signal-to-noise-ratio (PSNR) in terms of correlations with subjective quality evaluations [1]. It has been exciting to observe the new progress in both theoretical development and novel techniques on this multidisciplinary topic, which appears to be a converging point from a wide range of research directions and includes the following:

- signal and image processing
- computer vision
- visual psychophysics
- neural physiology
- information theory
- machine learning
- design of image acquisition, communication, and display systems.

While the field of objective IQA is still evolving quickly, and novel and better IQA methods will continue to emerge in the coming years, it is also interesting to discuss how we could make the best use of these tools in real-world applications. The purpose of this article is to provide an overview of the roles of IQA methods in these applications. We will start by a brief description of the current status of the IQA field, followed by discussions on benchmarking and monitoring applications of IQA measures. We will then discuss the applications of IQA measures in the design and optimization of advanced image processing algorithms and systems, where we perceive both great promises and major challenges. Finally, we will show how IQA measures could play important roles in an even more extended field of applications and provide a vision of the future.

Digital Object Identifier 10.1109/MSP.2011.942295
Date of publication: 1 November 2011

OBJECTIVE IMAGE QUALITY ASSESSMENT

Objective IQA measures aim to predict perceived image/video quality by human subjects, which are the ultimate receivers in most image processing applications. Depending on the availability of a pristine reference image that is presumed to have perfect quality, IQA measures may be classified into full-reference (FR), reduced-reference (RR), and no-reference (NR) methods. FR measures require full access to the reference image, while NR methods assume completely no access to the reference. RR methods provides a compromise in-between, where only partial information in the form of RR features extracted from the reference image is available in assessing the quality of the distorted image. IQA measures may also be categorized into application-specific or general-purpose methods. The former only applies to some specific applications where the types of distortions are often known and fixed, e.g., JPEG compression. The latter is employed in general applications, where one may encounter diverse types and levels of image distortions.

In the literature, a considerable number of IQA algorithms have been proposed, which exhibit substantial diversity in the methodologies being used. Meanwhile, they also share some common characteristics. In particular, all of them are rooted from certain knowledge in one or more of the following three categories (which, interestingly, constitute the basic building blocks in an information communication framework [1]):

1) knowledge about the image source, which can be either deterministic (when the reference image is fully available) or statistical (when certain statistical image models are employed)
2) knowledge about the distortion channel, which is often associated with some known facts about the specific distortion process that the images underwent, for example, blocking and blurring artifacts in JPEG compression, and blurring and ringing effects in wavelet-based image compression
3) knowledge about the receiver, i.e., the human visual system (HVS), where computational models originated from visual physiological and psychological studies play essential roles.

Until now, the area that has achieved the greatest success is FR IQA of gray-scale still images. Several algorithms, including the structural similarity index (SSIM) [2] and its derivatives, and the visual information fidelity (VIF) [3], significantly outperformed PSNR and MSE in a series of tests based on large-scale subject-rated independent image databases. There have also been notable success in the areas of video quality assessment (VQA) as well as RR and NR IQA, especially application-specific methods [4].

1053-5888/11/$26.00©2011IEEE

On the other hand, there is also an abundant menu of unresolved IQA problems left for future studies, including the following:

- General-purpose RR and NR IQA, where the types of image distortions are unavailable, is still at an immature stage.
- Methods for effective IQA of texture images are still lacking.
- There have not been good solutions for cross-dynamic range and cross-resolution IQA, where the reference image is available but has a different dynamic range of intensity levels or a different spatial or temporal resolution from the image being assessed.
- IQA for image signals with extended dimensions creates many challenging research problems, which include video, color, multispectrum, hyperspectrum, stereo, multiview, and three-dimensional (3-D) volume IQA.
- IQA algorithms that can be used for evaluating segmentation, halftoning, and fusion algorithms are lacking.
- In pattern recognition applications, effective IQA methods are missing that can assess how the recognition accuracy is affected by image distortions.
- In medical imaging applications, it is highly desirable to evaluate how image distortions affect the diagnostic values (rather than perceptual appeal) in images.
- In network visual communications, it is worth investigating how information regarding the communication channel conditions, such as channel fading characteristics and packet loss rate and delay could be utilized in the IQA process.
- In multimedia systems, visual quality may not be the only factor that affects the overall quality-of-experience (QoE) of users. Joint audio-visual quality assessment and joint quality assessment and visual discomfort evaluations are ongoing research topics. The complication of QoE assessment is raised to an even higher level in immersive multimedia environments such as panoramic 3-D displays.

In the past few years, there has been great effort in the research community to develop advanced IQA measures to solve the problems described above. For example, many recent projects carried out by the Video Quality Experts Group (VQEG) [12] are attempting to address these issues. Meanwhile, there are also many attempts to apply objective IQA measures for a wide variety of real-world applications, which will be our major focus in the next sections.

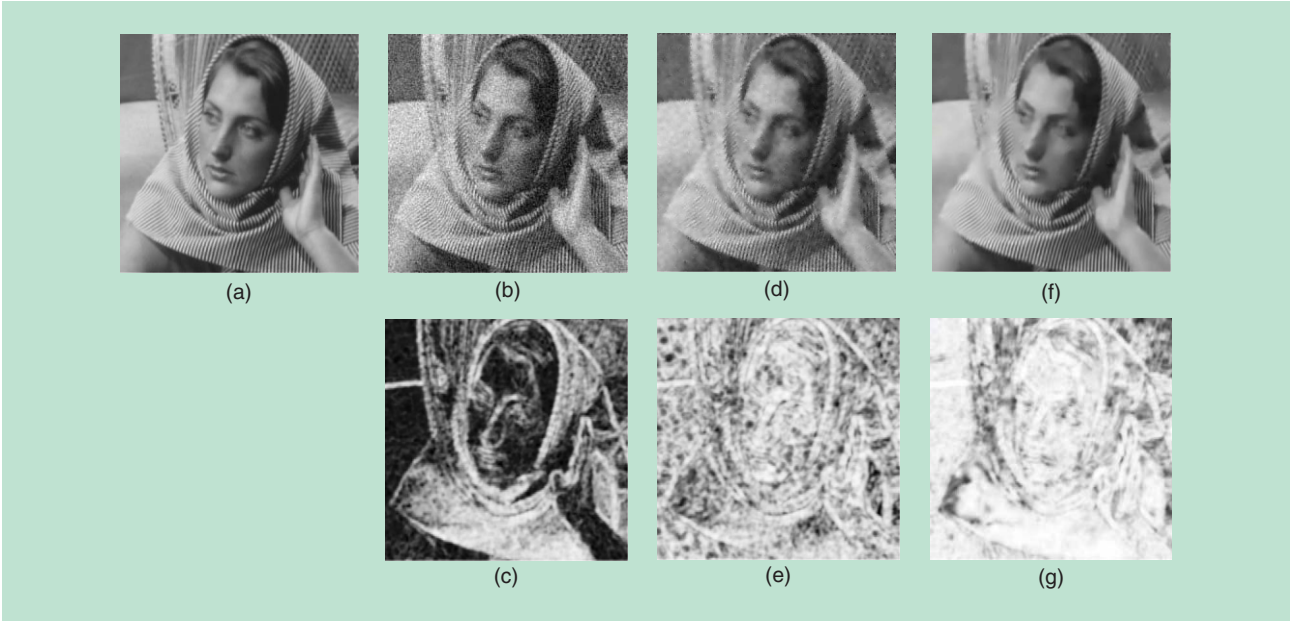
BENCHMARKING AND MONITORING APPLICATIONS

A direct application of IQA measures is to use them to benchmark image processing algorithms and systems. For instance, when multiple image denoising and restoration algorithms are available to recover images distorted by noise contamination and blur, a perceptual objective IQA measure can help pick the one that generates the best perceptual quality of the restored images. For another example, rate-distortion (RD) curves are often used to characterize the performance of image coding systems, where the RD function is defined as the distortion between the original and decoded images versus bit rate. A lower RD curve indicates a better image coder. Traditionally, MSE types of measures are employed to compute the distortion. If the role of MSE is replaced by a distortion function defined based on a perceptual IQA measure, then the RD curve could provide a perceptually more meaningful evaluation of the image coder.

IN MANY IMAGE PROCESSING ALGORITHMS, THERE ARE CERTAIN PARAMETERS THAT NEED TO BE DETERMINED BY USERS TO YIELD THE BEST RESULTS.

A useful feature of many IQA measures that is often overlooked by practitioners is that they not only create overall quality scores of distorted images, but also produce quality maps that indicate local quality variations over the image space. An example is given in Figure 1, where the original "Barbara" image (a) is contaminated by additive white Gaussian noise. Two denoising algorithms, spatially adaptive Wiener filtering (MATLAB Wiener2 function) and K singular-value-decomposition (KSVD) filtering [5], are employed to recover the original image from its noisy observation. The quality maps created by the SSIM index [2], a popular IQA measure, for the noisy image (b) and the denoised images (d) and (f) are given by (c), (e), and (g), respectively. These quality maps provide useful information in several aspects. First, despite the fact that noise is imposed uniformly over space, the perceptual quality varies significantly across the image. For example, the face region looks much noisier than the texture regions. These are clearly indicated by the quality map (c); Second, the quality maps help identify where in the image the denoisers yield the most improvement, and how one denoiser outperforms another. For instance, by comparing (e) and (g), we observe significant improvement of KSVD over Wiener filtering on the smooth regions as well as the stripe texture regions at the bottom part of the image. Third, the quality maps also indicate where the denoisers still need further improvement. For example, the textures in the upper-right region of the image are not well denoised by both algorithms.

In many image processing algorithms, there are certain parameters that need to be determined by users to yield the best results. This is often a difficult task for naive users as the best values may be image dependent. A good IQA measure could be a useful tool to help decide on these parameters automatically. For example, in [6], the Q-index, an NR sharpness and contrast measure, was used to automatically pick the parameters for image denoising algorithm. The idea may be extended further when multiple complementary algorithms (or multiple modes under the



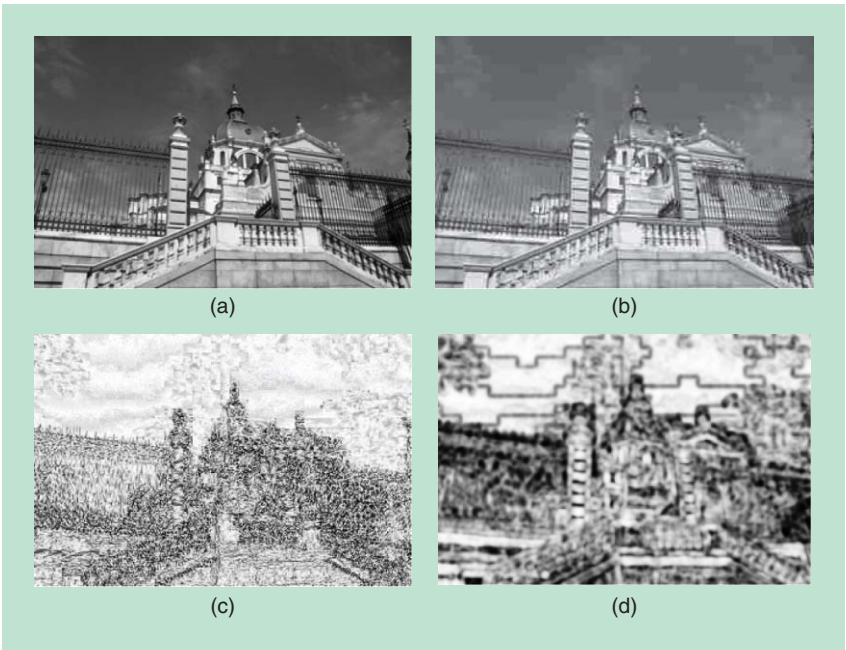
[FIG1] **Example of performance analysis using IQA measures and quality maps. An original image (a) is contaminated by noise and (b) denoised by two denoising algorithms, resulting in (d) and (f), respectively. The SSIM-based quality maps [2] of the noisy and denoised images are shown in (c), (e), and (g), respectively, where brighter indicates better local quality.**

same algorithm) are available for the same goal, for example, different coding modes in standard video compression systems. In such scenarios, an IQA measure can help select the right algorithm (mode), or to automatically switch between different algorithms (modes).

Objective IQA measures are particularly desirable in network visual communication applications for the purpose of quality-of-service (QoS) monitoring. Image and video content delivered over various wired and wireless networks inevitably suffer from visual quality degradations during lossy compression and transmission over error prone networks. It is imperative for the network service providers to monitor such quality degradations in real time, so as to optimize network resource allocations and maximally satisfy user expectations within certain cost constraints. It was shown that typical error criteria used in network design and testing, such as bit error rate (BER), do not correlate well with the quality of experience of network consumers [4]. Therefore, accurate and high-speed perceptual IQA measures can play important roles. Apparently, FR IQA methods are less useful here because the original video signals (typically with extremely high data rate) would not be available at the mid or end nodes in the network. NR methods are desired but are difficult to develop. This is mainly due to the complication of the types of distortions that could occur during video transmission in modern communication networks, where the distortions could be caused by a combination of lossy compression, network delay and packet loss, scaling in temporal and spatial resolution, scaling in bandwidth, spatial and/or temporal interpolation at the receiver, and various types of pre- and post-processing filtering (e.g., error concealment, deblocking filtering, and sharpening). RR IQA provides a useful compromise between FR and NR solutions, where RR features extracted from the original images are transmitted to the receiver end to evaluate the quality of the received distorted images. It was shown that with only a fairly low RR data rate, one may achieve impressive quality prediction accuracy close to competitive FR methods [7].

OBJECTIVE IQA MEASURES ARE PARTICULARLY DESIRABLE IN NETWORK VISUAL COMMUNICATION APPLICATIONS FOR THE PURPOSE OF QoS MONITORING.

The difficulty with RR based methods is how to transmit the RR features to the receiver. This typically requires a guaranteed ancillary channel, which may be expensive or unavailable. An interesting method to trace network image quality degradations without using an ancillary channel is to incorporate modern image watermarking techniques [8]. The idea is to hide a watermark image or a pseudo-random bit sequence inside the image being transmitted. The degradation of the watermark image or the error rate of the embedded bit sequence gauged at the receiver side can then be employed as an indicator of the quality degradation of the host image. The idea of quality-aware image provides another means to incorporate watermarking techniques [7], where RR features extracted from the original image are embedded into the same image as invisible hidden messages. When a distorted



[FIG2] An original image (a) is compressed by JPEG (b). The absolute error map and the SSIM quality map are shown in (c) and (d), respectively. In both maps, brighter indicates better local quality (or lower distortion).

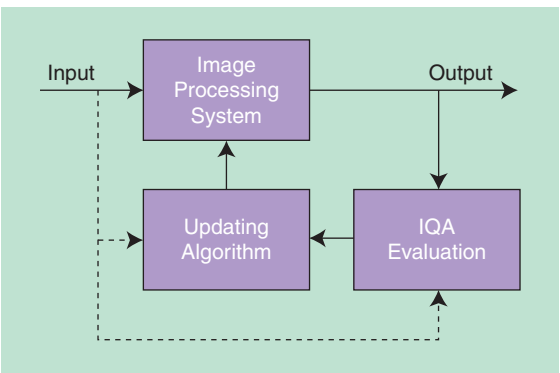
version of such a "quality-aware" image is received, users can decode the hidden RR features and use an RR IQA method to evaluate the quality of the distorted image. The advantages of watermarking-based methods are manifold. First, they do not require a separate data channel to transmit RR features or any other side information to the receiver. Second, they do not affect the conventional usage of the image content, because the data hiding process causes only invisible changes to the image. Third, compared with the approaches of including side information in image headers, they are more likely to survive image/video format conversion [7]. An additional and interesting benefit of quality-aware images is that they provide an opportunity for the end users to partially "repair" the received images using the decoded RR features. Such an idea of self-repairing image was demonstrated in [9] by matching certain statistical properties of the distorted image with those of the reference image (which are received as RR features). It was shown that this approach is quite effective for image deblurring [9].

DESIGN APPLICATIONS

The application scope of objective IQA measures is far beyond quality evaluation and algorithm comparison. In essence, any scientific design of image processing algorithms and systems would involve certain quality criterion, either explicitly or implicitly. If a good quality criterion is available, one may use it not only to assess the performance of these algorithms and systems, but also to optimize them so as to produce the best performance under this criterion.

Figure 2 demonstrates how a perceptual objective IQA measure could be useful in the context of image coding. An original image (a) is compressed using JPEG. Due to a limited bit budget, the resulting decompressed image (b) exhibits many highly visible distortions. In particular, the blocking artifacts in the sky can be clearly seen. The loss of details in the fence areas and the upper boundaries of the building is also obvious. Assume that some new bit budget is now available, and our goal is to decide on how to spend the new bits to enhance the image quality. Ideally, we would spend the bits at the locations that have the greatest potentials to improve the image quality. An IQA measure could assist us in identifying these locations. Figure 2(c) shows the absolute error map between (a) and (b), which is the first step in computing MSE and PSNR (as well as any l_p norm). Unfortunately, this error map provides us with wrong guidance, because it suggests that the inner parts of the building are where the largest distortions are located. By contrast, our visual observations are well consistent with the SSIM map (d) created by a perceptual IQA measure [2]. Realizing that most existing image coders are designed to optimize MSE/PSNR or similar criteria, the dramatic difference between the quality/error maps in (c) and (d) reveals the great potentials of perceptual image and video coding. Some recent work has shown great promises along this direction [4].

In the optimal design of image processing algorithms and systems, objective IQA measures may be employed in two different approaches. In the first approach, the core image processing module is kept unaltered, and the IQA measure is only used to create feedback control signals that help update the image processing module, likely in an iterative manner. This is illustrated in Figure 3, where depending on the application, either FR, RR, or NR IQA measures may be employed to create the feedback control signal. For example, in the case of image enhancement, an NR method may be employed and only the image created at the output end is needed for IQA



[FIG3] Diagram of IQA-based feedback optimization method.

computation. In image coding applications, an FR IQA measure could be used that requires both decoded image from the output end and the original reference image from the input (which is linked through the dashed line).

In the second approach of IQA-based optimal design, the objective IQA measure goes into the core of the image processing algorithm. To illustrate this, let us use the general image reconstruction problem as an example. Assume that there exists an original image X that is unknown to us. What is available is some distorted or partial information Y produced by applying an operator D on X: $Y = \mathrm{D}(X)$. Our goal is to design a reconstruction operator R, which, when applied to Y, yields a reconstructed image $\hat{X} = \mathrm{R}(Y)$, so that $\hat{X}$ is as close to X as possible. Depending on the operator D, this formulation could describe many practical problems. For example, when D denotes noise contamination, then this is a denoising problem. When D represents a downsampling operator, then it corresponds to an interpolation problem. Similarly, the same general framework could cover other problems such as image deblurring, decompression, inpainting, and reconstruction from compressed sensing data. Most of these problems are ill posed, in the sense that the solutions are not unique. To make the problem mathematically sound, one would need to define a cost function as the goal for minimization. For example, in a statistical approach, one treats X as a random variable associated with certain probability distribution and may define the optimization problem as

$$\hat{X}_{\mathrm{opt}} = \min_{\hat{X}} \mathbb{E}\{d(X, \hat{X})|Y\}, \qquad (1)$$

where $\hat{X}_{\mathrm{opt}}$ denotes the optimal solution, $\mathbb{E}$ represents the expectation operator, and d is an image distortion measure. The "standard" option for d is the MSE. To convert this to a perceptual optimization problem is straightforward—replacing d with a monotonically decreasing function with respect to a perceptual IQA measure.

Although the second approach for IQA-based optimal design looks appealing, when it comes to solving the optimization problem in (1), one often faces major difficulties. This is largely due to the lack of desirable mathematical properties in most perceptual IQA measures. To understand this, let us consider why the MSE is still the prevailing optimization criterion, regardless of the wide criticism on its poor correlation with perceptual image quality (as demonstrated in Figure 2). Indeed, the MSE is an ideal target for optimization [1]. It is based on a valid distance metric (l_2) that satisfies positive definiteness, symmetry, and triangular inequality properties. It is convex, differentiable, memoryless, and additive for independent sources of distortions. It is also energy preserving under orthogonal or unitary transformations [1]. Hardly any perceptual IQA measures with good quality prediction performance satisfies any of these properties. In [10], some initial attempts has been made to develop novel image distortion metrics that approximate the SSIM index while maintaining some of the desirable mathematical properties. It was shown that a valid distance metric exists that can very well approximate the SSIM index. In addition, the metric also possesses some useful convexity properties.

WITH THE FAST ADVANCES OF MEDICAL IMAGING TECHNOLOGIES, THE AMOUNT OF MEDICAL IMAGE DATA BEING ACQUIRED EVERY DAY HAS BEEN INCREASING DRAMATICALLY.

EXTENDED APPLICATIONS

In most of the IQA applications we discussed so far, the final outputs are images. Besides these, IQA measures may also be extended to an even broader range of applications where the outputs are interpretations or classification labels of images. Image-based pattern recognition is one such example, where the quality of images is often a critical factor that affects the accuracy of the recognition algorithms. For example, in biometrics, the purpose is to recognize humans or verify human identities based on one or more unique physiological characteristics of humans. Many biometric methods are based on images, including images of faces, fingerprints, palmprints, hand shapes, and handwritings. In practice, the acquisition process of these images may not be perfect, and thus the biometric systems may have to work under the conditions of noisy, distorted, or partially impaired images. In these application scenarios, it would be useful to know the level of quality degradations of these images and what recognition accuracy can be expected under such quality degradations. Different from traditional performance evaluation of IQA measures, here the IQA measures should be assessed and compared based on how they can predict the impact of image quality degradations on the final recognition performance, rather than the perceptual appealingness of the images. Once the image quality is estimated, some preprocessing procedure may be performed to enhance the quality of the image before the pattern recognition algorithm is applied. Another way of using the IQA results is to use them to help select between multiple recognition algorithms or to fuse the results from multiple algorithms, so as to improve the overall recognition performance. Such an approach has been successfully used in fingerprint verification systems [11].

With the fast advances of medical imaging technologies, the amount of medical image data being acquired every day has been increasing dramatically, largely surpassing the increase of available storage space. Efficiently storing, transmitting, and retrieving medical image information in large-scale databases has become a major challenge in hospitals and medical organizations. Lossy image compression provides a powerful means to reduce the data rate, but runs the risk of losing or altering

important diagnostic information contained in medical images. It is therefore important to provide specific objective IQA measures that can help maximize the level of compression, but without affecting the diagnostic value of medical images. Moreover, many modern medical imaging devices acquire images with much higher dynamic range of intensity levels than what can be appropriately shown on standard dynamic range displays. Therefore, it is desirable to employ those IQA measures that can provide meaningful quality evaluations of the images after dynamic range compression. Furthermore, both data rate and dynamic range compression of medical images should be optimized for the IQA measures specifically designed for medical applications.

OUTLOOK

We have discussed the application aspects of modern objective IQA methods. Rather than providing an exhaustive survey of all applications, we have emphasized on the great potentials of IQA applications, provided instructive examples, and also discussed the main challenges. In the future, it is expected that the development and application sides of objective IQA measures will mutually benefit each other. On one hand, more accurate and more efficient IQA measures will certainly enhance their applicability in real-world applications. On the other hand, new challenges arising from real applications (e.g., desired mathematical properties for optimization purposes) will impact the new development of future IQA measures.

AUTHOR

Zhou Wang (zhouwang@ieee.org) is an associate professor in the Department of Electrical and Computer Engineering, University of Waterloo, Canada.

REFERENCES

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Mag.,* vol. 26, pp. 98–117, Jan. 2009.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing,* vol. 13, pp. 600–612, Apr. 2004.

[3] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Processing,* vol. 15, pp. 430–444, Feb. 2006.

[4] H. R. Wu and K. R. Rao, Ed., *Digital Video Image Quality and Perceptual Coding.* Boca Raton: CRC, 2005.

[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Processing,* vol. 15, pp. 3736–3745, Dec. 2006.

[6] X. Zhu and P. Milanfar, "Automatic parameter selection for denoising algorithms using a no-reference measure of image content," *IEEE Trans. Image Processing,* vol. 19, pp. 3116–3132, Dec. 2010.

[7] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Processing*, vol. 15, pp. 1680–1689, June 2006.

[8] M. C. Q. Farias, S. K. Mitra, M. Carli, and A. Neri, "A comparison between an objective quality measure and the mean annoyance values of watermarked videos," in *Proc. IEEE Int. Conf. Image Processing,* Rochester, MN, Sept. 2002, pp. 469–472.

[9] A. Rehman and Z. Wang, "Reduced-reference SSIM estimation," in *Proc. IEEE Int. Conf. Image Processing,* Hong Kong, China, Sept. 2010, pp. 289–292.

[10] D. Brunet, E. R. Vrscay, and Z. Wang, "A class of image metrics based on the structural similarity quality index," in *Proc. Int. Conf. Image Analysis and Recognition (Lect. Notes Comput. Sci.)*, vol. 6753, Burnaby, BC, Canada, June 2011, pp. 100–110.

[11] J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia, and A. K. Jain, "Incorporating image quality in multi-algorithm fingerprint verification," *Lect. Notes Comput. Sci.,* vol. 3832, pp. 213–220, 2005.

[12] Video Quality Experts Group Web site [Online]. Available: www.vqeg.org

[SP]

fed into a spatial light modulator, a device that can modulate light spatially in amplitude and phase. "You could get all that information and display it in 3-D and it can actually be in real time," he says.

While most existing telepresence systems are geared toward conferencing applications, Peyghambarian feels that real-time holography has the potential to drive the technology into a wider range of fields. "Benefits include 3-D social networking, 3-D remote surgery, and 3-D collaborative research," he says. "The advantage of our technology is that it can continuously read and replace data, so you can use it in an magnetic resonance imaging or computer assisted tomography scan system that would provide the information it gathered in 3-D to doctors." The technology could, for example, help surgeons performing brain surgery or other types of delicate operations. "They could use that [technology] to see the information as they do the operation," Peyghambarian says.

John Apostolopoulos, director of the Mobile and Immersive Experience Lab at Hewlett-Packard (HP) Laboratories in Palo Alto, California, believes that signal processing will be vital to overcoming many of the challenges telepresence researchers currently face. "This includes video and audio capture, noise reduction, compression, transmission over a packet network, packet-loss concealment, multichannel echo cancellation, efficient signal-processing algorithms for multicore and GPU systems and so on," he says. "I believe that advances in signal processing will continue to be central to improving telepresence in the future."

None of these improvements will come too soon for Microsoft's Zhang, who admits that he has a personal interest in seeing sophisticated telepresence systems becoming commonplace. "I have frequent phone calls with my parents and family members in China as well as my research collaborators at Microsoft Research Asia in Beijing," he says. "Telephony is a great invention, but leaves much more to be desired compared with a face-to-face meeting."

[SP]

lecture NOTES

J. David Osés del Campo, Fernando Cruz–Roldán,
Manuel Blanco–Velasco, and Sergio L. Netto

A Single Matrix Representation for General Digital Filter Structures

A matrix representation of a general single-input single-output (SISO) digital filter structure is addressed, proposing a single matrix that stores the complete description of the filter in a very compact and functional format. The proposed matrix contains the structural information corresponding to the block diagram (BD) connections and, at the same time, it can be seen as a valid computational algorithm to implement the filter in the time domain. With this matrix, the gap between the signal flow graph (SFG) and a bit-true implementation of the filter can be considerably reduced. Finally, simple methods to derive the matrix representation from/onto the block diagram and the state-space (SS) mapping are also described and illustrated with some examples.

RELEVANCE

New digital filter structures are continuously proposed, presenting strategies to improve the behavior of the systems under, for instance, finite word length conditions [1]–[4]. In this context, and considering the increasing interest in mapping applications onto field-programmable gate arrays and other signal processors, it is very important to have a compact mathematical representation of the system that specifies the exact order in which the computations must be performed, allowing for the simulation of different alternatives and the selection of the best-suited realization for one's particular purposes. To reduce the time required to implement the filter, it is convenient to complement the information of classical BDs or SFGs with a mathematical description, in terms of matrices, which is closer to a valid ready-to-use computational algorithm.

Different representations can be found in the associated literature for describing, in matrix form, the equations corresponding to a general BD with N nodes. In here, a new matricial representation is proposed with the following interesting characteristics: it is very compact; it preserves all information from the original filter structure; it can be readily transformed onto the BD, SFG, or the SS descriptions; and it can be easily mapped on a computable algorithm implementing the desired digital filter.

PROBLEM STATEMENT

Among the currently known digital filter descriptions, one of the most compact and interesting [5], proposed by Crochiere and Oppenheim, has the general form

$$\mathbf{y}[n] = \mathbf{x}[n] + \mathbf{F}_{\mathbf{c}}^{\mathrm{T}}\mathbf{y}[n] + \mathbf{F}_{\mathbf{d}}^{\mathrm{T}}\mathbf{y}[n-1], \tag{1}$$

where

- $\mathbf{y}[n]$ is an $N \times 1$ column vector of the node signal values at instant n
- $\mathbf{x}[n]$ is an $N \times 1$ column vector of the input signal values at instant n
- $\mathbf{F}_{\mathbf{c}}^{\mathrm{T}}$ is an $N \times N$ matrix of coefficient branches
- $\mathbf{F}_{\mathbf{d}}^{\mathrm{T}}$ is an $N \times N$ matrix of coefficient-delay branches.

Alternative descriptions can be found in [6] and [7]. In model (1), the constant matrices $\mathbf{F}_{\mathbf{c}}^{\mathrm{T}}$ and $\mathbf{F}_{\mathbf{d}}^{\mathrm{T}}$ do not completely describe the system. To obtain a general model, completely characterized by constant matrices, the input vector $\mathbf{x}[n]$ can be written as $\mathbf{E}x[n]$, where for an N–node SISO network, $\mathbf{E}$ is a constant $N \times 1$ column matrix indicating the node where the input $x[n]$ is applied, and a possible scaling factor k_x. For simplicity, we shall employ the following notation for the general description based only on three constant matrices:

$$\mathbf{w}[n] = \mathbf{E}x[n] + \mathbf{F}\mathbf{w}[n] + \mathbf{G}\mathbf{w}[n-1], \tag{2}$$

where $\mathbf{F}$ and $\mathbf{G}$ are equivalent to $\mathbf{F}_{\mathbf{c}}^{\mathrm{T}}$ and $\mathbf{F}_{\mathbf{d}}^{\mathrm{T}}$, respectively, in (1). The vector $\mathbf{w}[n]$ contains the node signal values. By convention, and without loss of generality, we assume that the last element of $\mathbf{w}[n]$ corresponds to the filter output $y[n]$.

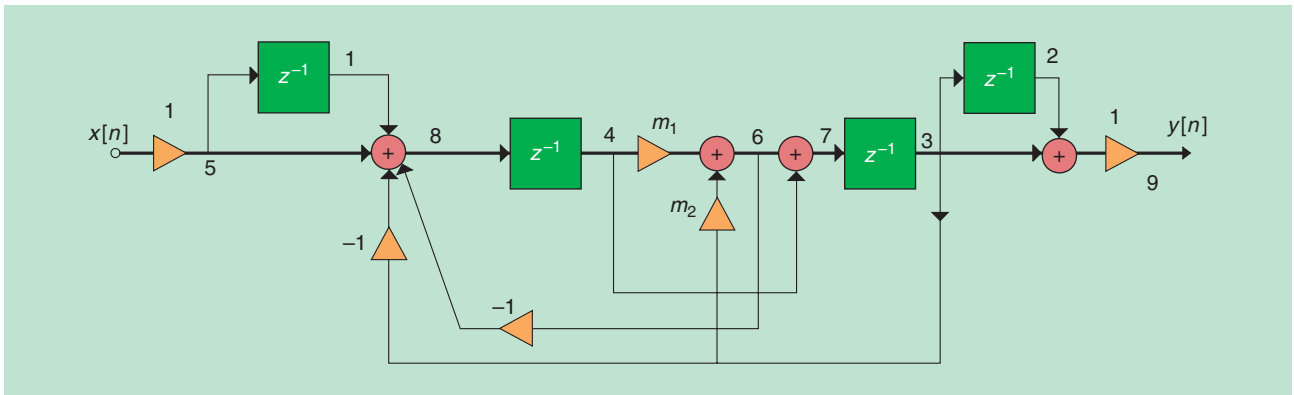
Let us consider as an example the filter of Figure 1 (which corresponds with Figure 2 of [8]), where the nodes in the BD have been previously labeled to obtain a valid and computable set of equations.

We can describe this structure by means of the following system of equations:

$$\mathcal{S}: \begin{cases} w_1[n] = w_5[n-1] \\ w_2[n] = w_3[n-1] \\ w_3[n] = w_7[n-1] \\ w_4[n] = w_8[n-1] \\ w_5[n] = x[n] \\ w_6[n] = w_3[n]\cdot m_2 + w_4[n]\cdot m_1 \\ w_7[n] = w_4[n] + w_6[n] \\ w_8[n] = w_1[n] - w_3[n] + w_5[n] - w_6[n] \\ y[n] = w_2[n] + w_3[n] \end{cases}. \tag{3}$$

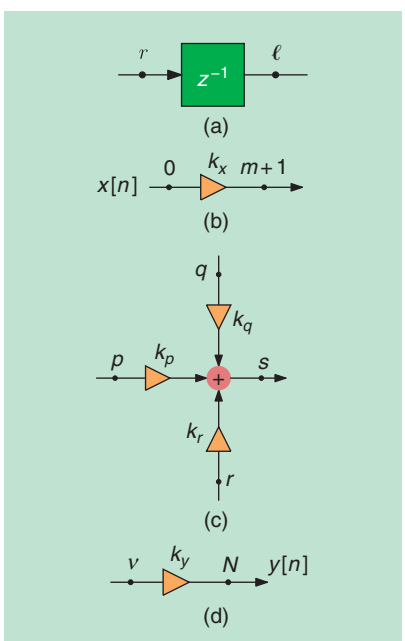
and the corresponding matrix representation, according to model (2), becomes (4) as seen in the box on the next page.

Digital Object Identifier 10.1109/MSP.2011.942837
Date of publication: 1 November 2011



[FIG1] **Modified CGIC lowpass filter from [8].**

$$\begin{bmatrix} w_1[n] \\ w_2[n] \\ w_3[n] \\ w_4[n] \\ w_5[n] \\ w_6[n] \\ w_7[n] \\ w_8[n] \\ y[n] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} x[n] + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_2 & m_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1[n] \\ w_2[n] \\ w_3[n] \\ w_4[n] \\ w_5[n] \\ w_6[n] \\ w_7[n] \\ w_8[n] \\ y[n] \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1[n-1] \\ w_2[n-1] \\ w_3[n-1] \\ w_4[n-1] \\ w_5[n-1] \\ w_6[n-1] \\ w_7[n-1] \\ w_8[n-1] \\ y[n-1] \end{bmatrix}. \tag{4}$$



[FIG2] **Graphical representations of (a) delay element corresponding to each row of $\hat{\mathbf{G}}$. (b) Input branch of $\hat{\mathbf{E}}$. (c) Adder element corresponding to several rows of $\hat{\mathbf{F}}$ with the same first element. (d) Output branch corresponding to last row of $\hat{\mathbf{F}}$.**

Taking the z-transform of (2), we get

$$\mathbf{W}(z) = \mathbf{E}\mathbf{X}(z) + \mathbf{F}\mathbf{W}(z) + \mathbf{G}\mathbf{W}(z)z^{-1}. \tag{5}$$

Therefore, the response of the different nodes considered in $\mathbf{w}[n]$ to the input $x[n]$ can be expressed, in the z-domain, as $\mathbf{W}(z) = \mathbf{T}(z)\mathbf{X}(z)$, with the so-called transfer-function matrix given by

$$\mathbf{T}(z) = \begin{bmatrix} t_1(z) \\ t_2(z) \\ \vdots \\ t_{N-1}(z) \\ t_N(z) \end{bmatrix} = [\mathbf{I} - \mathbf{F} - z^{-1}\mathbf{G}]^{-1}\mathbf{E}, \tag{6}$$

where $t_i(z)$ denotes the transfer function from the input to the ith system node. From the assumption that the last model node corresponds to the system output, the filter transfer function is given by

$$H(z) = t_N(z) = \frac{Y(z)}{X(z)}. \tag{7}$$

SOLUTION

Typically, matrices $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ in model (2) are quite sparse. Hence, they can be represented by more compact forms, thus avoiding unnecessary space for zero-valued elements. In addition, we can also take advantage of this sparsity to minimize the amount of computation needed to solve the associated equations.

For example, the `sparse` function of MATLAB generates the following storage organization: (`row`, `column`) `entry`. Using this notation, matrices $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ corresponding to the filter depicted in Figure 1 can be compactly expressed as shown in Table 1.

Looking at this sparse representation of matrices, it is clear that we can compress and summarize the complete filter model by combining the three constant matrices, with slight changes in the row orders and a zero introduced in $\mathbf{E}$, to form one single partitioned matrix, which we shall call $\mathbf{M}$, as follows:

$$\mathbf{M} = \begin{bmatrix} \hat{\mathbf{G}} \\ \hat{\mathbf{E}} \\ \hat{\mathbf{F}} \end{bmatrix}, \tag{8}$$

where we use the modified notation $\hat{\mathbf{G}}$ and $\hat{\mathbf{F}}$ to indicate that the rows of $\mathbf{G}$ and $\mathbf{F}$ have been sorted in increasing order of their first elements. In addition, $\hat{\mathbf{E}}$ indicates a change to zero of the second entry of $\mathbf{E}$. In short, we have that

- $\hat{\mathbf{G}}$ is an $m \times 3$ matrix, m being the number of delays, of rows of the form $[\ell \ \mathrm{r}\ 1]$, denoting a delay block from node r to node ℓ, which corresponds to the operation $w_\ell[n] = w_r[n-1]$ and the circuit element depicted in Figure 2(a).
- $\hat{\mathbf{E}}$ is a 1×3 vector corresponding to the input branch, with the second entry always set to zero to allow us to identify the input node. By convention, and without loss of generality, we may consider that the input comes from node 0 to node $(m+1)$ with a scaling factor of k_x, such that $w_{m+1} = x[n] \cdot k_x$, as indicated in Figure 2(b).
- $\hat{\mathbf{F}}$ is a three-column matrix with the information of all adder-multiplier branches properly ordered. Repeated entries s in the first column, with nodes p, q, and r in the second column and respective gains k_p, k_q, and k_r for different rows, correspond to the relationship $w_s[n] = w_p[n] \cdot k_p + w_q[n] \cdot k_q + w_r[n] \cdot k_r$, as illustrated in Figure 2(c). If an element in the first column of $\hat{\mathbf{F}}$ appears only once, no addition is performed and a single multiplication gives the value at the corresponding output node. Once again, by convention and without loss of generality, the last row corresponds to the filter output, whose graphical representation is depicted in Figure 2(d).

The first column of $\mathbf{M}$ is an ordered list of natural numbers, where the numbers greater than $(m+1)$ can be repeated. The zero entry in the second column of $\mathbf{M}$ corresponds to the input signal and allows us to identify and separate, in a simple manner, the information corresponding to $\hat{\mathbf{G}}$, $\hat{\mathbf{E}}$, and $\hat{\mathbf{F}}$. The last entry, N, of the first column identifies the output node of the filter. In other words, the general matrix $\mathbf{M}$ is in the form

[TABLE 1] MATLAB COMPACT REPRESENTATION OF SPARSE MATRICES E, F, AND G IN (4).

E		F		G	
(5,1)	1	(8,1)	1	(2,3)	1
		(9,2)	1	(1,5)	1
		(6,3)	m_2	(3,7)	1
		(8,3)	-1	(4,8)	1
		(9,3)	1		
		(6,4)	m_1		
		(7,4)	1		
		(8,5)	1		
		(7,6)	1		
		(8,6)	-1		

$$\mathbf{M} = \begin{bmatrix} 1 & \cdot & 1 \\ 2 & \cdot & 1 \\ 3 & \cdot & 1 \\ \vdots & \vdots & \vdots \\ m & \cdot & 1 \\ \hline m+1 & 0 & k_x \\ \hline m+2 & \cdot & \cdot \\ m+2 & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ m+s & \cdot & \cdot \\ m+s & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ N & \cdot & k_y \end{bmatrix} \begin{matrix} \} \\ \\ \\ \\ \\ \} \\ \\ \\ \\ \\ \\ \\ \\ \\ \} \end{matrix} \begin{matrix} \text{Delays} \\ \\ \\ \\ \\ \text{Input} \\ \\ \\ \\ \text{Adders and} \\ \text{Multipliers} \\ \\ \\ \\ \\ \text{Output.} \end{matrix} \tag{9}$$

Using this newly proposed notation, all three matrices $\mathbf{E}$, $\mathbf{F}$, and $\mathbf{G}$ (usually large and sparse) in model (2) are replaced by a single and compact matrix $\mathbf{M}$ containing the same information. In our particular example, for the filter depicted in Figure 1, we have

$$\hat{\mathbf{G}} = \begin{bmatrix} 1 & 5 & 1 \\ 2 & 3 & 1 \\ 3 & 7 & 1 \\ 4 & 8 & 1 \end{bmatrix}, \quad \hat{\mathbf{E}} = [5 \quad 0 \quad 1],$$

$$\hat{\mathbf{F}} = \begin{bmatrix} 6 & 3 & m_2 \\ 6 & 4 & m_1 \\ 7 & 4 & 1 \\ 7 & 6 & 1 \\ 8 & 1 & 1 \\ 8 & 3 & -1 \\ 8 & 5 & 1 \\ 8 & 6 & -1 \\ 9 & 2 & 1 \\ 9 & 3 & 1 \end{bmatrix}. \tag{10}$$

NODE ORDERING

Now we describe a simple procedure to obtain the matrix $\mathbf{M}$ from the block diagram, guaranteeing the computability of the resulting system of equations.

Different algorithms can be found for developing a correct node-ordering sequence, leading to a computable system of difference equations. In [5], a formal procedure based on the node precedence relations of the digital network is explained. Another approach, based on a three-step algorithm appears in [7]. Usually, the difference equations characterizing the filter can be put in a computable order by simple inspection, guaranteeing that the signal at a particular node does not depend on the signal of a node whose output is yet to be determined.

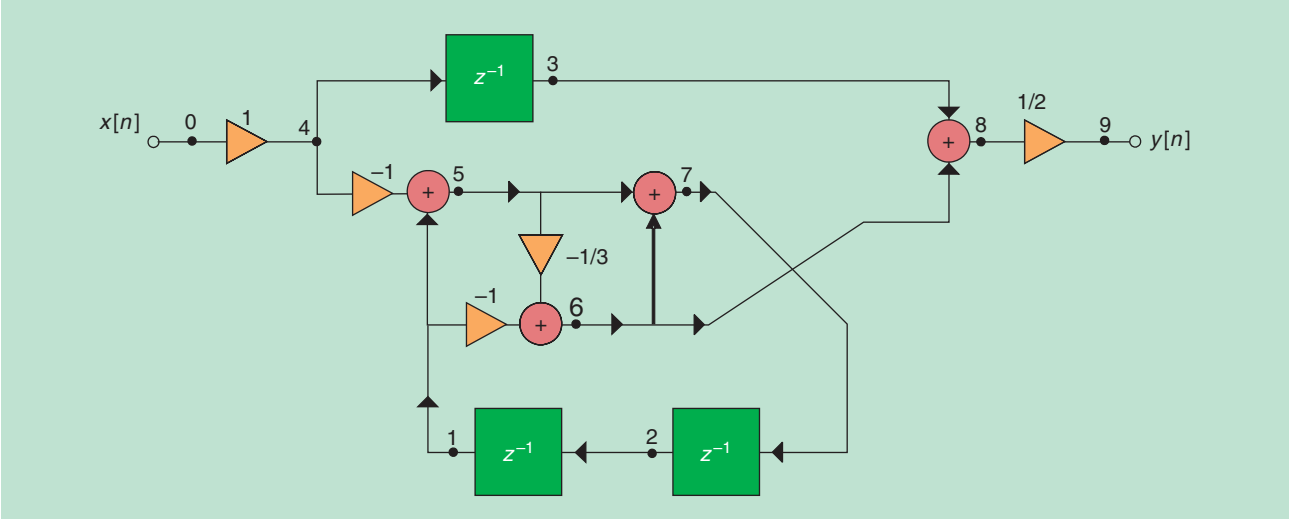
To organize the node-labeling process in a simple and efficient way to obtain matrix $\mathbf{M}$, we propose the following strategy for numbering the N nodes of a general network containing m unit delay blocks:

i) The first nodes $(1, 2, m)$ will be the output of the m delay elements. When several delays are connected in series or the output of one delay block is connected to the input of another one, we shall enumerate the corresponding nodes in increasing order beginning at the outermost output. As examples, see nodes 2 and 1 in Figure 3, or nodes 3 and 2 in Figure 1, respectively.

ii) The input signal, $x[n]$, is connected from node zero to node $(m+1)$.

iii) We shall enumerate the rest of the nodes taking into account the fact that a new node can only be labeled if it depends on the signals previously determined. In general, these nodes will correspond to the output of the adders. The output of the filter, $y[n]$, will be the last node enumerated.

To illustrate the proposed process, we have labeled the nodes of the bireciprocal-lattice wave digital filter [9] presented in Figure 3, where $m = 3$. Initially, and by simple inspection, we can apply steps i) and ii) to number nodes $0, 1, \ldots, (m+1)$. As step iii) indicates, we continue enumerating the output of the adders, starting by the adders whose



[FIG3] **Example of node ordering result for wave digital filter [9].**

inputs have already been labeled, and, finally, the last node is assigned to the output.

Once the nodes have been properly labeled, matrix $\mathbf{M}$ can be determined, as described above, following the established node order. In this case, for the digital filter in Figure 3, we get

$$\mathbf{M} = \left[\begin{array}{ccc|c|cccccccc} 1 & 2 & 3 & 4 & 5 & 5 & 6 & 6 & 7 & 7 & 8 & 9 \\ 2 & 7 & 4 & 0 & 4 & 1 & 5 & 1 & 5 & 6 & 3 & 8 \\ 1 & 1 & 1 & 1 & -1 & 1 & -\tfrac{1}{3} & -1 & 1 & 1 & 1 & \tfrac{1}{2} \end{array}\right]^{\mathrm{T}}. \tag{11}$$

Computability for the compact model can be checked by simple inspection of the submatrices $\hat{\mathbf{G}}$ and $\hat{\mathbf{F}}$ of matrix $\mathbf{M}$. Taking into account the fact that delay registers are updated with values calculated in the previous iteration (or initial conditions), the first entry of any row of matrix $\hat{\mathbf{G}}$ must be smaller than the second element of the same row. On the other hand, since a node can only be computed if all the required data are available, the first entry of any row of matrix $\hat{\mathbf{F}}$ must be greater than the second element of the same row.

Matrix $\mathbf{M}$ in (11) corresponds to the sparse model (12) shown in the box at the bottom of the page. In this sparse representation, computability requires matrix $\mathbf{F}$ to be a lower-triangular matrix.

REVERSE MODELING AND SET OF EQUATIONS

Starting with the model matrix $\mathbf{M}$, one can readily obtain the BD or the SFG of the associated digital filter using the graphical representations seen in Figure 2. For that purpose, we must first identify the corresponding $\hat{\mathbf{G}}$, $\hat{\mathbf{E}}$, and $\hat{\mathbf{F}}$ submatrices by locating the null entry in the second column of $\mathbf{M}$. Hence, each row in $\hat{\mathbf{G}}$ represents a delay element, whereas all rows with identical entry in the first column of $\hat{\mathbf{F}}$ define a multiply-and-add element; finally, the input and output branches are respectively characterized by $\hat{\mathbf{E}}$ and the last row of $\hat{\mathbf{F}}$.

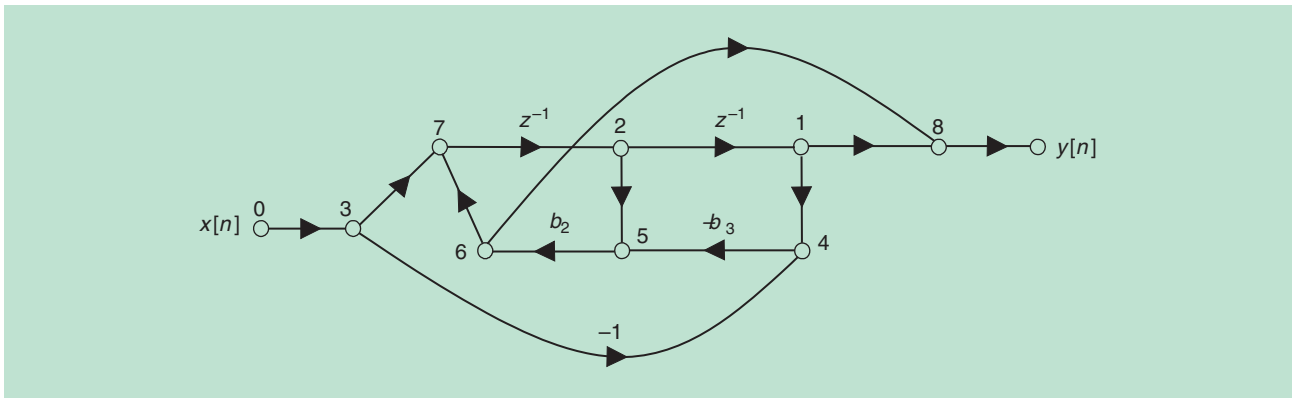
Following this strategy for matrix $\mathbf{M}$ determined by (10), one can readily obtain the digital filter represented in Figure 1. Alternatively, for the compact model

$$\mathbf{M} = \left[\begin{array}{cc|c|cccccccc} 1 & 2 & 3 & 4 & 4 & 5 & 5 & 6 & 7 & 7 & 8 & 8 \\ 2 & 7 & 0 & 1 & 3 & 2 & 4 & 5 & 3 & 6 & 1 & 6 \\ 1 & 1 & 1 & 1 & 1 & 1 & -b_3 & b_2 & 1 & 1 & 1 & -1 \end{array}\right]^{\mathrm{T}}, \tag{13}$$

the SFG [10] shown in Figure 4 results, where the node numbers have been assigned as specified by the contents of $\mathbf{M}$.

In addition, the proposed model also allows us to obtain, in a simple and direct manner, the corresponding set of difference equations without the necessity of determining any graphical representation for the associated filter. In fact, following the same reasoning as above, the

$$\begin{bmatrix} w_1[n] \\ w_2[n] \\ w_3[n] \\ w_4[n] \\ w_5[n] \\ w_6[n] \\ w_7[n] \\ w_8[n] \\ y[n] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} x[n] + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & -\tfrac{1}{3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \tfrac{1}{2} & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1[n] \\ w_2[n] \\ w_3[n] \\ w_4[n] \\ w_5[n] \\ w_6[n] \\ w_7[n] \\ w_8[n] \\ y[n] \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1[n-1] \\ w_2[n-1] \\ w_3[n-1] \\ w_4[n-1] \\ w_5[n-1] \\ w_6[n-1] \\ w_7[n-1] \\ w_8[n-1] \\ y[n-1] \end{bmatrix}. \tag{12}$$



[FIG4] SFG corresponding to the compact model provided in (13).

model matrix $\mathbf{M}$ in (13) can be easily mapped onto the set of equations

$$\mathcal{S}:\begin{cases} w_1[n] = w_2[n-1] \\ w_2[n] = w_7[n-1] \\ w_3[n] = x[n] \\ w_4[n] = w_1[n] + w_3[n] \\ w_5[n] = w_2[n] + w_4[n] \cdot (-b_3) \\ w_6[n] = w_5[n] \cdot b_2 \\ w_7[n] = w_3[n] + w_6[n] \\ y[n] = w_1[n] - w_6[n] \end{cases} \tag{14}$$

In the Appendix, a generalization of the classical `filter` function of MATLAB is presented. In this new function, the information of the filter structure is specified by the user incorporating the matrix M as the first parameter to be introduced. The main advantage of this generalization lies in the flexibility afforded when finite word-length effects are studied; these effects can be easily simulated using the quantize function. Taking into account that the different arithmetical operations to be performed are univocally specified in matrix M, the exact order in which the quantization errors appear, are propagated and finally accumulated in a certain register, can be analyzed. This framework facilitates the comparison of different structures.

STATE-SPACE MAPPING

Now we focus our attention on the relationship between the proposed model and the SS representation [6], [7].

It must be mentioned that the compact matrix $\mathbf{M}$ retains all information regarding the original system architecture, allowing us to transform this model to/from any alternative representations such as the BD, SFG, set of difference equations, or the sparse matrix model, as discussed above. Meanwhile, the SS description is a distinct compact representation of a given system, including the input-output relationship as well as its internal operation, that does now allow a reverse mapping onto, for instance, the original SFG.

Despite this major difference, the proposed model can be readily mapped onto the SS description through the following algorithm:

i) Associate each of the m delay output nodes, corresponding to the first-column entry in each row of $\hat{\mathbf{G}}$, to a system state $u_j[n+1]$, for $j = 1, 2, \ldots, m$.

ii) Identify the input $x[n]$ and output $y[n]$ nodes from $\hat{\mathbf{E}}$ and the last row of $\hat{\mathbf{F}}$, respectively.

APPENDIX

```
function [y,cf]=filt_M(M,x,ci)
% [y,cf]=filt_M(M,x,ci) filters data in vector x with the
% filter described by matrix M.
% INPUT PARAMETERS
% M → Matrix M.
% x → Input data.
% ci → Initial conditions of the delay blocks.
% OUTPUT PARAMETERS
% y → Output data.
% cf → Final state of the delay blocks.
L=max(M(:,1));
w=zeros(L,1);
m=find(M(:,2)==0)-1; % number of delay blocks
w([M(1:m,2)])=ci; % initial conditions
for n=1:length(x)
  for i=1:m
    w(i)=M(i,3)*w([M(i,2)]); % updating delay blocks
  end
  w(m+1)=M(m+1,3)*x(n); % input
  w(m+2:end)=0;
  for i=m+2:size(M,1) % rest of nodes
    r=w(M(i,2))*M(i,3); % multiply
    w(M(i,1))=w(M(i,1))+r; % addition
  end
  y(n)=w(end); % output (last node)
end
cf=w([M(1:m,2)]); % final conditions of delay blocks
```

iii) Solve partially the set $\mathcal{S}$ of difference of equations corresponding to matrix $\mathbf{M}$, writing each state $u_j[n+1]$ as a function of $u_j[n]$, $x[n]$ and $y[n]$. By doing so, one must eliminate dependence to any variable $w_i[n]$ not associated to a system state.

iv) Write the difference equations found in Step iii) in the form

$$\mathcal{U}: \begin{cases} \mathbf{u}[n+1] = \mathbf{A}\mathbf{u}[n] + \mathbf{B}x[n] \\ y[n] = \mathbf{C}\mathbf{u}[n] + \mathbf{D}x[n] \end{cases}, \quad (15)$$

where

- $\mathbf{u}[n] = [u_1[n]\ u_2[n]\ \ldots\ u_m[n]]^{\mathrm{T}}$ is the state vector at time n.
- $\mathbf{A}$ is the $m \times m$ system matrix, $\mathbf{B}$ is an $m \times 1$ column vector, $\mathbf{C}$ is a $1 \times m$ row vector, and $\mathbf{D}$ is a scalar.

Following this approach, we can readily obtain the SS description for each of the digital filters depicted in Figure 1:

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & m_2 & 1+m_1 \\ 1 & 0 & -m_2-1 & -m_1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

$$\mathbf{C}^{\mathrm{T}} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \mathbf{D} = 0; \quad (16)$$

Figure 3:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ -\frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 \\ -\frac{2}{3} \\ 1 \end{bmatrix},$$

$$\mathbf{C}^{\mathrm{T}} = \begin{bmatrix} -\frac{2}{3} \\ 0 \\ \frac{1}{2} \end{bmatrix}, \mathbf{D} = \frac{1}{6}; \quad (17)$$

and Figure 4:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -b_2b_3 & b_2 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 \\ 1-b_2b_3 \end{bmatrix},$$

$$\mathbf{C}^{\mathrm{T}} = \begin{bmatrix} 1+b_2b_3 \\ -b_2 \end{bmatrix}, \mathbf{D} = b_2b_3. \quad (18)$$

CONCLUSIONS

A new compact model based on a single matrix $\mathbf{M}$ that describes the input-output relationship, as well as all internal information, of a general digital filter has been presented, and its application widely illustrated. This matrix $\mathbf{M}$ complements the block diagram representation, preventing possible errors during the signal-flow graph analysis and, at the same time, saving all internal data in a very compact manner. Moreover, matrix $\mathbf{M}$ can be interpreted as a simple alternative for the computable set of difference equations that describes the digital-filter evolution in time, shortening considerably the time required to implement the system. Finally, a simple node-ordering sequence oriented to facilitate the construction of matrix $\mathbf{M}$ has also been proposed and relationship to the SS representation was explicited, emphasizing the positive aspects of the proposed model.

ACKNOWLEDGMENTS

This work was partially supported by the Spanish Ministry of Science and Innovation through project TEC2009-08133, the Ministry of Education through project PHB2008-0017, and by the Brazilian Research Agency CAPES through project 205/09.

AUTHORS

J. David Osés del Campo (doses@ics.upm.es) (doses@ics.upm.es) is an associate professor at the Universidad Politécnica de Madrid, Spain.

Fernando Cruz-Roldán (fernando.cruz@uah.es) is a professor at the Universidad de Alcalá (UAH), Spain.

Manuel Blanco-Velasco (manuel.blanco@uah.es) is an associate professor at UAH, Spain.

Sergio L. Netto (sergioln@lps.ufrj.br) is an associate professor at Federal University of Rio de Janeiro, Brazil.

REFERENCES

[1] K. Uesaka and M. Kawamata, "Evolutionary synthesis of digital filter structures using genetic programming," *IEEE Trans. Circuits Syst. II*, vol. 50, no. 12, pp. 977–983, Dec. 2003.

[2] T.-B. Deng and Y. Nakagawa, "SVD-based design and new structures for variable fractional-delay digital filters," *IEEE Trans. Signal Processing*, vol. 52, no. 9, pp. 2513–2527, Sept. 2004.

[3] R. Lehto, T. Saramäki, and O. Vainio, "Synthesis of narrowband linear-phase FIR filters with a piecewise-polynomial impulse response," *IEEE Trans. Circuits Syst. I*, vol. 54, no. 10, pp. 2262–2276, Oct. 2007.

[4] G. Li, J. Chu, and J. Wu, "A matrix factorization-based structure for digital filters," *IEEE Trans. Signal Processing*, vol. 55, no. 10, pp. 5108–5112, Oct. 2007.

[5] R. E. Crochiere and A. V. Oppenheim, "Analysis of linear digital networks," *Proc. IEEE*, vol. 63, no. 4, pp. 581–595, Apr. 1975.

[6] S. K. Mitra, *Digital Signal Processing. A Computer Based Approach*, 3rd ed. New York: McGraw-Hill, 2006.

[7] P. S. R. Diniz, E. A. B. da Silva, and S. L. Netto, *Digital Signal Processing. System Analysis and Design, 2nd ed.* Cambridge, U.K.: Cambridge Univ. Press, 2010.

[8] G. Paul and A. Pal, "High speed power efficient CGIC digital filters for VLSI applications," in *Proc. 2006 Annu. IEEE India Conf.*, Sept. 15–17, 2006, pp. 1–4.

[9] L. Gazsi, "Explicit formulas for lattice wave digital filters," *IEEE Trans. Circuits Syst.*, vol. 32, no. 1, pp. 68–88, Jan. 1985.

[10] S. Mitra and K. Hirano, "Digital all-pass networks," *IEEE Trans. Circuits Syst.*, vol. 21, no. 5, pp. 688–700, Sept. 1974. SP

evaluates their discusses existing metrics, including perceptually based ones, computed either on 3-D data or on 2-D projections, and evaluates their correlation performance with existing subjective studies.

Study Group 12 (SG12) of the Telecommunication Standardization Section of the International Telecommunication Union (ITU-T) has been involved for many years in standardizing methods for multimedia quality assessment, both subjective and objective. "Multimedia Quality Assessment Standards in ITU-T SG12," by Coverdale et al., gives an overview of existing and emerging SG12 standards, with a special focus on models that predict quality on the basis of parameters and bit stream information available during network planning and monitoring phases.

We hope that this special issue has reached its objective of providing researchers and professionals in the field of multimedia signal processing with timely articles addressing not only the latest advances in the evaluation and assessment of multimedia quality, but also trends and challenges, which in turn provides a solid basis for further progress in this exciting and dynamic field. Enjoy reading! SP

dsp TIPS&TRICKS

Abhishek Seth and Woon-Seng Gan

Fixed-Point Square Roots Using *L*-b Truncation

"DSP Tips and Tricks" introduces practical design and implementation signal processing algorithms that you may wish to incorporate into your designs. We welcome readers to submit their contributions. Contact Associate Editors Rick Lyons (R.Lyons@ieee.org) or Clay Turner (clay@claysturner.com).

In this article, we describe several techniques to reduce computational workload of the Newton-Raphson (NR)-based fixed-point square rooting method. Using the described techniques, the computational workload of NR methods can be reduced at the expense of memory. These new techniques outperform existing fixed-point square rooting methods both in terms of results accuracy and computational efficiency.

COMPUTING SQUARE ROOTS

Square root (SQRT) operations are used in a number of applications, like spectrum analysis, audio signal processing, digital communication and three-dimensional (3-D) graphics. Fixed-point digital signal processors (DSPs) are widely used in the aforementioned areas and many others. This makes fixed-point SQRT an important arithmetic operator. There are many methods described in literature to find the SQRT of a number [1]. Based on their structures, some of them are more suitable for hardware implementation, while others are more suited for software implementation on digital signal processors with a hardware multiplier. In practice, only a limited amount of computational resources can be assigned to a square rooting routine. Choosing a fixed-point square rooting method requires a tradeoff between its computational load and bit accuracy.

Digital Object Identifier 10.1109/MSP.2011.942297
Date of publication: 1 November 2011

DIRECT NEWTON-RAPHSON METHOD

The direct Newton-Raphson (DNR) method for computing SQRTs [2] is an iterative technique to find SQRT of a number x using

$$y_{k+1} = y_k + \frac{1}{2\sqrt{x}}(x - y_k^2),\ k = 0,1.....,\tag{1}$$

where y_{k+1} is the estimated value of SQRT of x obtained after $(k+1)$ iterations.

Here we will use the following notation:

$$\text{SQRT}(x) = \sqrt{x},$$
$$\text{and} \qquad \text{ISQRT}(x)/2 = \frac{1}{2\sqrt{x}}.\tag{2}$$

DNR requires an initial approximation for SQRT(x) and ISQRT$(x)/2$. The initial approximation for SQRT(x) is represented by y_0 and this approximation is improved after each iteration. Initial approximations of SQRT(x) and ISQRT$(x)/2$ can be obtained using either polynomial expansions (PEs) or look-up tables (LUTs) [1]. The block diagram of a general DNR using two iterations to find the SQRT of a 16-b fixed-point number is shown in Figure 1. In our notation, the subscripted number 16 in x_{16} and $y_{0,16}$, for example, in Figure 1, means that those samples are 16 b in width.

As shown in Figure 1, a single iteration of DNR requires two multiplications and two additions. For the SQRT(x) and ISQRT$(x)/2$ initialization in (1), either PE or LUT can be used. PE increases the computational load, whereas LUT increases memory consumption. A variation of DNR, known as the nonlinear infinite impulse response filter (NLIIRF) method, is described in [3]. The NLIIRF method, which uses PE for SQRT(x) initialization and an LUT for ISQRT$(x)/2$ initialization, is described next.

NLIIRF METHOD

The NLIIRF method is an iterative technique to find the SQRT of a number [3]. The input range of the NLIIRF method, described in [3], is stated as $0.25 \le x < 1$. Any value of x lying outside this range needs to be normalized to this range. The iterative computation used to determine the SQRT of x is given by

$$y_{k+1} = y_k + \beta(x - y_k^2),\ k = 0,1,2,\ \ldots\tag{3}$$

The nonlinear difference expression in (3) can be implemented by an IIR filter—hence the term "NLIIRF method." In (3), y_k is the approximation for SQRT(x) at the kth iteration. However, during the first iteration of (3), the starting approximation for SQRT(x) (stated as y_0) needs to be found using a linear PE. In addition, β can be considered as an approximation of ISQRT$(x)/2$ and determined by different methods. In [3], β is stored in a LUT. In [4], the authors suggested that a linear PE of the form of

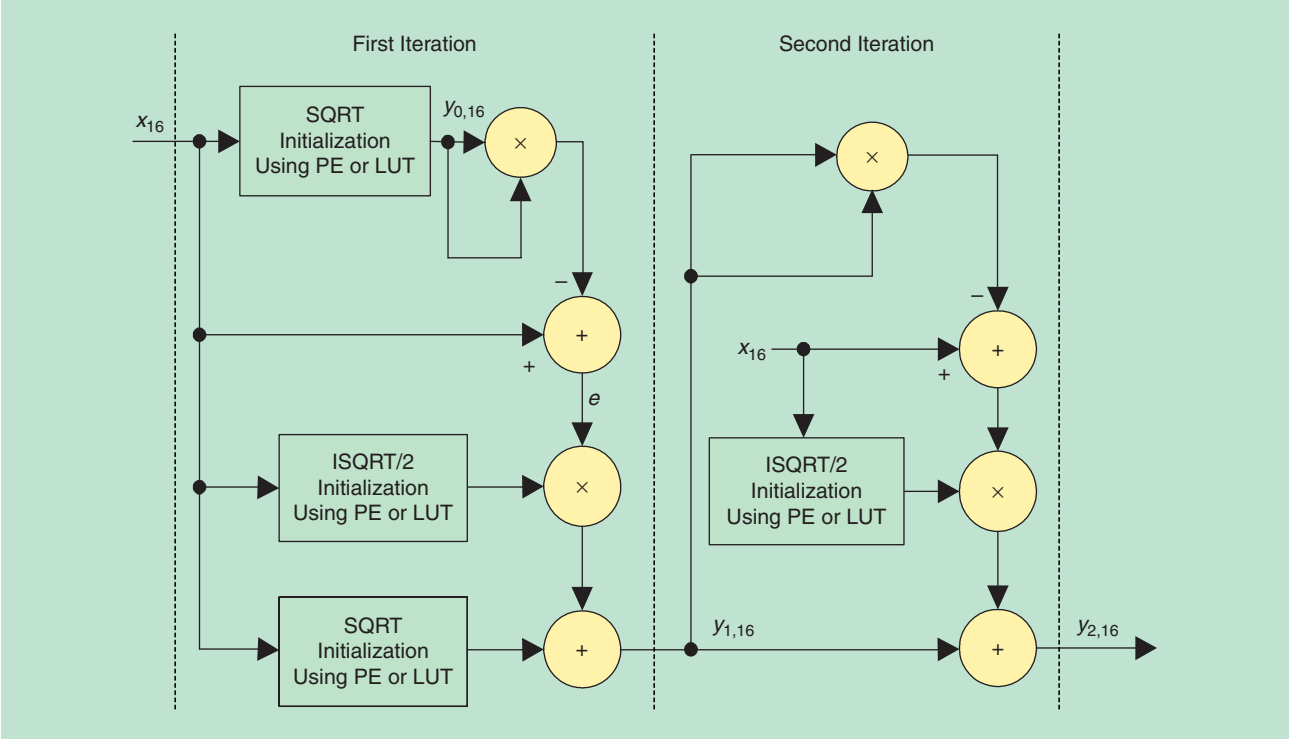
$$\beta = -0.61951x + 1.0688,\tag{4}$$

or a quadratic PE of the form of

$$\beta = 0.763x^2 - 1.5688x + 1.314\tag{5}$$

can be used. The NLIIRF method can also be depicted by Figure 1.

While the NLIIRF SQRT method is viable, we next propose a specialized



[FIG1] **Block diagram of a general DNR algorithm.**

DNR method that reduces the computational workload below that of the NLIIRF method.

PROPOSED $\mathrm{DNR_T}(n)$ METHOD

For a 16-b fixed-point processor, the NLIIRF method can produce highly accurate results with two iterations (using five multiplications and seven additions). We therefore aim to formulate a method that can generate equal to or more accurate results on a 16-b fixed-point processor with fewer than five multiplications and seven additions. We call our proposed SQRT method "$\mathrm{DNR_T}(n)$," where the symbol "(n)" in $\mathrm{DNR_T}(n)$ denotes the number of multiplications used by the algorithm.

The proposed $\mathrm{DNR_T}(n)$ method uses LUTs to store initial approximate values of both $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$. A direct advantage of using LUTs over PEs is that it saves a number of additions and multiplications at the expense of memory needed for the LUTs.

The input range for the $\mathrm{DNR_T}(n)$ spans from $0.25 \le x < 1$ as described in [4]. To generate the $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ LUTs, the interval $0.25 \le x < 1$ is divided into subintervals with equal width of $2^{-(L-1)}$ $(L < 16)$. This gives $0.75 \cdot 2^{(L-1)}$ different values, denoted as x_L, obtained by rounding x_{16} to L b (1 b for sign and $(L-1)$ b for magnitude). The $\mathrm{SQRT}(x_L)$ and $\mathrm{ISQRT}(x_L)/2$ values are computed for each x_L, rounded to 16 b, and stored in the $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ LUTs respectively. So the $\mathrm{SQRT}(x)$ LUT and $\mathrm{ISQRT}(x)/2$ LUTs each have $0.75 \cdot 2^{(L-1)}$ entries, where each entry is 16 b in width. The index of each LUT memory location, integer I_{DX}, is in the range $0 \le I_{\mathrm{DX}} \le 0.75 \cdot 2^{(L-1)} - 1$ and is computed using

$$I_{\mathrm{DX}} = (x_L - 0.25) \cdot 2^{(L-1)} \qquad (6)$$

requiring one subtraction operation and a multiply implemented by an arithmetic left shift. The same I_{DX} value is used for $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ LUT. Sample x_{16} is rounded to x_L by adding 2^{-L} to $x_{16}$and truncating the lower $(16-L)$ b of the result. Therefore, including (6), one addition and one subtraction are required to generate index for LUTs.

In the first iteration of Figure 1, a $\mathrm{SQRT}(x)$ LUT value is squared to produce an x_L data sample. That x_L is subsequently subtracted from x_{16} to generate error term e as

$$e = x_{16} - x_L. \qquad (7)$$

We can compute e without using (7). If x_{16} is truncated to L b, the truncation error term, denoted by e_{T}, is obtained by extracting the least significant $(16-L)$ bits of x_{16}. Our desired error term e is then calculated from e_{T} as

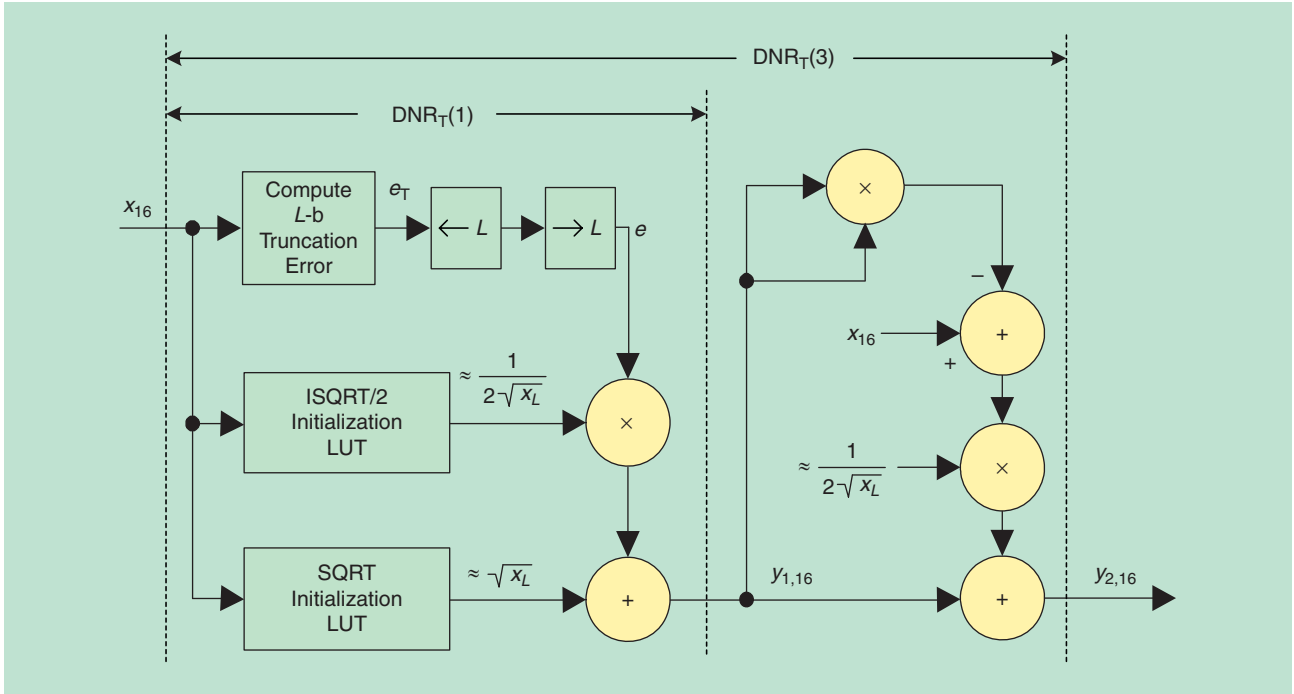
$$e = (e_{\mathrm{T}} \leftarrow L) \rightarrow L, \qquad (8)$$

where '$\leftarrow L$' ('$\rightarrow L$') means an arithmetic left (right) shift by L b. Therefore, our first trick is to replace the squaring and subtraction in the first iteration of Figure 1 with L-b truncation and arithmetic shifts as shown in Figure 2.

Next we show a two-multiply $\mathrm{DNR_T}(2)$ scheme that approximates the $y_{2,16}$ output in Figure 2.

A $\mathrm{DNR_T}(2)$ METHOD

Due to the range of data values at various nodes in Figure 2, we can make a few reasonable assumptions about that



[FIG2] **Block diagram of DNR$_T$(1) and DNR$_T$(3).**

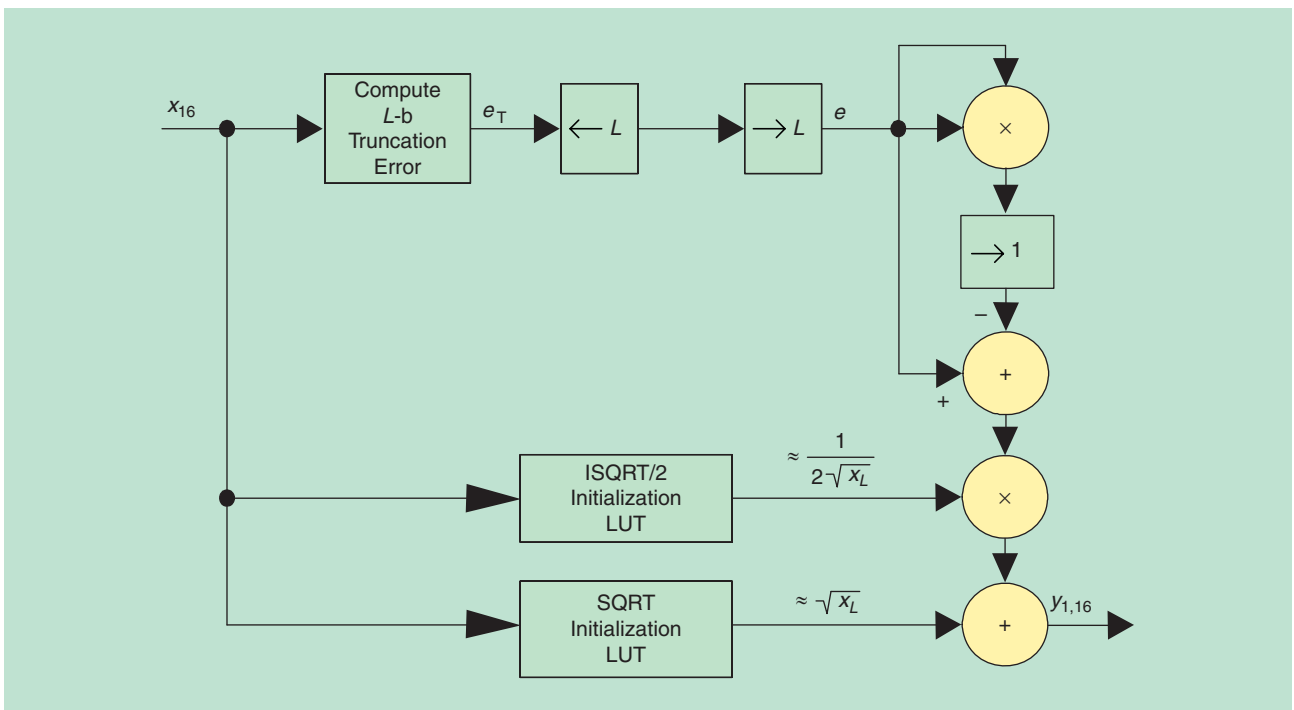
data and approximate the $y_{2,16}$ output using only one processing iteration. We call that single-iteration scheme the "DNR$_T$(2) method" and depict it in Figure 3.

The $y_{2,16}$ output in Figure 2 is approximated by the $y_{1,16}$ output in Figure 3, requiring only two multiplications, using

$$y_{1,16} \approx y_{2,16} \approx \mathrm{LUT}_{\mathrm{SQRT}}(i) + \mathrm{LUT}_{\mathrm{ISQRT/2}}(i) \cdot (e - e^2/2). \quad (9)$$

The derivation of (9) is provided in the Appendix.

If the $e^2/2$ term in (9) is neglected, we see that (9) represents the DNR$_T$(1) method in Figure 2. Next we show an improved DNR$_T(n)$ scheme that uses enhanced precision (EP) LUTs.



[FIG3] **Block diagram of DNRT$_T$(2).**

APPENDIX

The derivation of the $\mathrm{DNR}_T(2)$ processing expression in (9) proceeds as follows: to keep the notation simple we refer to the output of the $\mathrm{ISQRT}(x)/2$ LUT, $\mathrm{LUT}_{\mathrm{ISQRT}/2}(i)$, as I and refer to the output of the $\mathrm{SQRT}(x)$ LUT, $\mathrm{LUT}_{\mathrm{SQRT}}(i)$, as S. Referring to Figure 2, denoting the difference between x_{16} and x_L by e, we write

$$e = x_{16} - x_L, \tag{A1}$$

where the output of first iteration is given by

$$y_{1,16} = S + e \cdot I. \tag{A2}$$

Subsequently, the $y_{2,16}$ output of the second iteration can be written as

$$y_{2,16} = [x_{16} - (S + e \cdot I)^2] \cdot I + S + e \cdot I$$

$$= S + I \cdot (x_{16} - S^2 - e^2 \cdot I^2) + e \cdot I - 2 \cdot S \cdot I \cdot e \cdot I. \tag{A3}$$

Because $S \cdot I \approx 1/2$, the last two terms in (A3) cancel each other, and the $x_{16} - S^2$ terms can be replaced by e, giving us

$$y_{2,16} \approx S + I \cdot (e - e^2 \cdot I^2), \tag{A4}$$

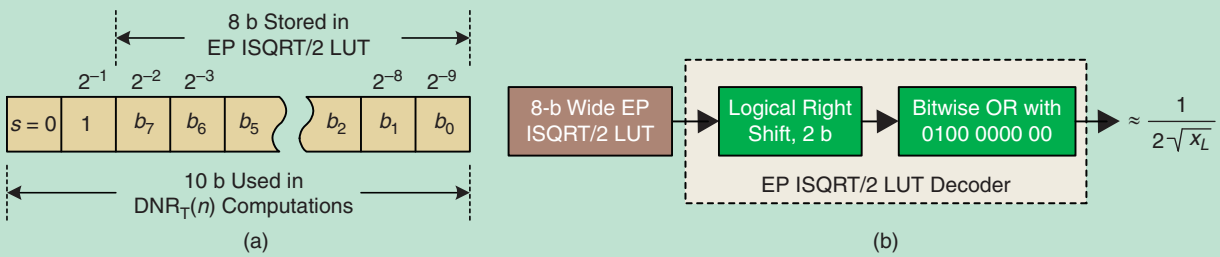
or

$$y_{2,16} \approx \mathrm{LUT}_{\mathrm{SQRT}}(i) + \mathrm{LUT}_{\mathrm{ISQRT}/2}(i) \cdot [e - e^2 \cdot (\mathrm{LUT}_{\mathrm{ISQRT}/2}(i))^2]. \tag{A5}$$

Our empirical studies have shown that $(\mathrm{LUT}_{\mathrm{ISQRT}/2}(i))^2$ can replaced by the factor 1/2 without inducing significant error in the desired results. As such, this gives a single-iteration approximation of

$$y_{1,16} \approx y_{2,16} \approx \mathrm{LUT}_{\mathrm{SQRT}}(i) + \mathrm{LUT}_{\mathrm{ISQRT}/2}(i) \cdot (e - e^2/2)$$

as presented in (9).



[FIG4] **Parts (a) and (b) show generating EP ISQRT(X)/2 LUT.**

ENHANCED PRECISION LUTs

Because the $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ values are always positive, the sign bit of the binary representations of those values is always zero. Now, for the x_{16} input interval $0.25 \le x_{16} < 1$, the contents of both LUTs are in the range of $0.5 \le$ table value < 1. So the most significant bit (MSB) (bit next to sign bit) is always one, for each entry of the LUTs. Our next trick is to make use of these facts to generate EP $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ LUTs. Figure 4(a) shows how an EP $\mathrm{ISQRT}(x)/2$ LUT value having 10-b precision can be stored in an 8-b LUT memory location. When that $\mathrm{ISQRT}(x)/2$ LUT value is accessed for a computation, the 8-b value is converted to an EP 10-b value as shown in Figure 4(b).

Similarly, we can generate a 16-b EP $\mathrm{SQRT}(x)$ LUT. Our simulation results show that we do not need to generate 16-b wide an EP $\mathrm{ISQRT}(x)/2$ LUT because an 8-b wide EP $\mathrm{ISQRT}(x)/2$LUT can generate sufficiently accurate results. This saves 25% of LUT memory.

Table 1 compares the performance of 16-b wide normal LUTs and EP LUTs. In Table 1, a listed value of L is the L used for L-b rounding. Those L values are a consequence of the chosen $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ LUT subinterval widths. For example, if the $\mathrm{SQRT}(x)$ and $\mathrm{ISQRT}(x)/2$ LUT subinterval width is 2^{-4} then $L = (1 + 4) = 5$. We interpret $(1+4)$ to be one sign bit and four magnitude bits (because processing is in Q1.15 binary arithmetic), so in this case x_{16} is rounded to 5 b.

We see from Table 1 that using EP LUTs not only reduces memory requirements but also generates more accurate results than using normal LUTs.

[TABLE 1] **NORMAL AND EP LUT SIZE AND PERFORMANCE COMPARISON.**

$\mathrm{DNR}_T(n)$	SUBINTERVAL WIDTH, L	TOTAL EP LUT SIZE (BYTES)		% OF OUTPUT SAMPLES WITH BIT PRECISION > 15	
		NORMAL LUT	EP LUT	NORMAL LUT	EP LUT
$\mathrm{DNR}_T(1)$	2^{-7}, 8	384	288	99.74	100
$\mathrm{DNR}_T(1)$	2^{-6}, 7	192	144	93.81	98.65
$\mathrm{DNR}_T(2)$	2^{-7}, 8	384	288	99.91	100
$\mathrm{DNR}_T(2)$	2^{-6}, 7	192	144	99.41	99.67
$\mathrm{DNR}_T(3)$	2^{-5}, 6	96	72	99.99	99.98
$\mathrm{DNR}_T(3)$	2^{-4}, 5	48	36	98.10	98.22

PROCESSING RESULTS COMPARISON

The above NLIIRF and $\mathrm{DNR}_T(n)$ methods are simulated for 8-b and 16-b fixed-point processors using MATLAB. The

[TABLE 2] COMPARISON OF VARIOUS FIXED-POINT SQUARE ROOT METHODS ON 8-b AND 16-b FIXED-POINT PROCESSORS.

	MULTIPLICATIONS		ADDITIONS		TOTAL LUT SIZE (BYTES)		% OF OUTPUT SAMPLES WITH BIT PRECISION > N–1	
METHOD	8 b	16 b	8 b	16 b	8 b	16 b	N = 8 b	N = 16 b
$DNR_T(1)$	1	1	3	3	12	288/144	100	100/98.65
$DNR_T(2)$	NOT APPLICABLE	2	NOT APPLICABLE	4	NOT APPLICABLE	288/144	NOT APPLICABLE	100/99.67
$DNR_T(3)$	NOT APPLICABLE	3	NOT APPLICABLE	5	NOT APPLICABLE	72/36	NOT APPLICABLE	99.98/98.22
NLIIRF (LUT)	3	5	5	7	15	30	100	99.3
NLIIRF (QUADRATIC PE)	6	8	5	7	7	14	100	97.3
NLIIRF (LINEAR PE)	4	6	6	6	6	12	87.50	40.19

fixed-point SQRT output is compared with the reference of a double-precision (DP) floating-point (FP) SQRT output, y generated from MATLAB, to evaluate the accuracy error. The negative $\log_2$ of the magnitude of error values is used to calculate bit precision for each output sample, as

$$\text{Bit precision of } y_{k,N} = -\log_2(|y - y_{k,N}|). \tag{10}$$

For 8-b and 16-b fixed-point processors, error values are generated for all values of x that can be represented by 8 and 16 b, respectively, lying in the interval [.25, 1). For N-b (eight or 16) fixed-point processors, number of additions, multiplications, bytes required, and percentage of output with bit precision $> N-1$ b are shown in Table 2. The L values for the $DNR_T(n)$ configurations in Table 2 are same as those in Table 1.

For both 8-b and 16-b fixed-point precision, NLIIRF (LUT) method performs better than NLIIRF (linear PE) and NLIIRF (quadratic PE) methods. Even though it uses fewer multiplications and fewer LUT bytes, the 8-b fixed-point $DNR_T(1)$ achieves the same degree of accuracy as the NLIIRF (LUT) method. Since bit accuracy > 7 b can be achieved with $DNR_T(1)$ for all output samples, $DNR_T(2)$ and $DNR_T(3)$ are not required for 8-b fixed-point precision. For 16-b fixed-point numbers, $DNR_T(1)$, $DNR_T(2)$, and $DNR_T(3)$ are used.

As shown in Table 2, for 16-b fixed-point precision, $DNR_T(1)$ with a total LUT (SQRT(x) LUT plus ISQRT(x)/2 LUT) size of 144 B ($L = 7$) produces 98.65% of output samples with bit accuracy > 15 b. This percentage value increases to 100% when the total LUT size is increased to 288 B ($L = 8$). Similar observations can be made for $DNR_T(2)$ and $DNR_T(3)$. For 16-b fixed-point precision, $DNR_T(n)$ for $n = 1$, 2, and 3, either outperforms or generate comparable results as compared to the NLIIRF(LUT). For 16-b fixed-point processor, $DNR_T(n)$ requires fewer number of multiplications and additions as compared to the NLIIRF(LUT). But on the downside, it needs longer LUTs. However, $DNR_T(n)$ provides the flexibility of decreasing the computational load at the expense of memory without compromising the performance. However, this flexibility is not available in the NLIIRF method.

CONCLUSIONS

This article presented a square rooting method suitable for 8-b and 16-b fixed-point implementation. $DNR_T(n)$ outperforms the NLIIRF method and its variants in terms of the number of operations (additions and multiplications) and accuracy. But $DNR_T(n)$ needs more memory to store LUTs. Because current fixed-point processors come with internal memories ranging up to megabytes, the memory requirements of $DNR_T(n)$ can be considered negligible for both 8-b and 16-b platforms. In this regard, $DNR_T(n)$ turns out to be a highly accurate and computationally cheap square rooting method for both 8-b and 16-b fixed-point processors. Moreover, the various $DNR_T(n)$ methods provide options for reducing computational load at the expense of memory. This scalability can be exploited in various situations making $DNR_T(n)$ a versatile algorithm. A Web site providing MATLAB m-files for NLIIRF and $DNR_T(n)$ are provided for the reader's reference [5].

ACKNOWLEDGMENT

The authors would like to acknowledge Rick Lyons for his fruitful discussions, which led to the improvement of this article. His insightful comments and suggestions formed the basis of some of the interesting ideas presented in the article.

AUTHORS

Abhishek Seth (aseth@ntu.edu.sg) is pursuing his M.Eng. degree (research) in the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore.

Woon-Seng Gan (ewsgan@ntu.edu.sg) is an associate professor in the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore.

REFERENCES

[1] P. Montuschi and M. Mezzalama, "Survey of square rooting algorithms," *Proc. Inst. Elect. Eng.*, vol. 137, pp. 31–40, Jan. 1990.

[2] C. Ramamoorthy, J. Goodman, and K. Kim, "Some properties of iterative square-rooting methods using high-speed multiplication," *IEEE Trans. Comput.*, vol. C-21, pp. 837–847, Aug. 1972.

[3] N. Mikami, M. Kobayashi, and Y. Yokoyama, "A new DSP-oriented algorithm for calculation of square root using a nonlinear digital filter," *IEEE Trans. Signal Processing*, vol. 40, pp. 1663–1669, July 1992.

[4] M. Allie and R. Lyons, "A root of less evil," *IEEE Signal Processing Mag.*, vol. 22, pp.93–96, Mar. 2005.

[5] A. Seth and W. Seng Gan. (2011). Matlab Square Root Files. [Online]. Available: http://www3.ntu.edu.sg/home/aseth/

[SP]

[exploratory DSP]

Manuel Richey and Hossein Saiedian

Compressed Two's Complement Data Formats Provide Greater Dynamic Range and Improved Noise Performance

In this article, we present and analyze a new family of fixed-point data formats for signal processing applications. These formats represent compressed two's complement data formats where compression on the sign bit is undertaken on a sample by sample basis. The extra room provided by sign-bit compression is utilized to retain more bits of precision for each individual sample. Compressed two's complement data formats are shown to provide greater dynamic range and improved noise performance over traditional fixed-point data formats such as sign-magnitude, offset binary, and traditional two's complement. Traditional two's complement is shown to be a member of the compressed two's complement family where the compression factor (CF) is equal to one. The dynamic range of a compressed two's complement data format is shown to approach the dynamic range of a non-compressed data format raised to the power of the CF. Improved performance for digital signal processing (DSP) applications such as digital filters and transforms is presented for specific instances of this family.

INTRODUCTION

The binary bits that make up an information stream are typically not all of equal importance. This inequality has fostered numerous data compression techniques, many of which focus on removing the least important bits while maintaining the essence of the information stream. Data compression is usually applied to information files such as text and pictures to reduce the amount of space required for storage. It is applied to information streams such as video and audio to decrease the required bandwidth needed to send information from one point to another. In this article, we present another use for data compression that improves the dynamic range and noise performance of discrete time signals.

Two's complement is the data format typically used to represent and operate on digital samples in a fixed-point DSP system because it provides the same numeric precision as other fixed-point data formats but is easier to implement in digital hardware [2]. In a two's complement fixed-point data format, the decimal point is set at a fixed position relative to the individual bits. As numbers become smaller, the sign bit fills in the unused digits at the most significant positions in the binary format. The primary assumption for this format is that the smallest digits are the least important, and they are typically rounded to reduce algorithmic noise and fit in the binary word width of the chosen two's complement format.

FOR SIGNAL PROCESSING APPLICATIONS, THE LEAST IMPORTANT BITS IN A TWO'S COMPLEMENT NUMBER ARE THE LEADING ONES AND ZEROS, NOT THE LEAST SIGNIFICANT BITS.

For signal processing applications, the least important bits in a two's complement number are the leading ones and zeros, not the least significant bits (LSBs). A number only needs one sign bit. All of the other leading digits are used to identify the distance between the significant digits and the decimal point. These leading ones and zeros can be efficiently compressed and the resulting space can be used for additional numeric precision. Unlike file-based compression techniques, this compression is performed on a sample-by-sample basis.

Figure 1 provides an example. Suppose a set of fixed-point data values of 12 b in width [Figure 1(a)] must fit into an 8-b data format. Typically, the entire data set is scaled such that the largest number fills the entire eight most significant bits (MSBs) [automatic gain control (AGC)]. Then the entire data set is rounded from 12 b to 8 b as shown in Figure 1(b) for a typical value. The negative effect of this approach is that up to 1/2 LSB of noise has been injected into each data sample by rounding. Furthermore, some of the

(a)	A 12-b Fixed-Point Number	000000101001
(b)	Rounded to an 8-b Two's Complement Number	00000011
(c)	Ideal Compressed Two's Complement Number	00101001

[FIG1] Data rounding from 12 b to 8 b of a typical fixed-point number (Red = sign bits, Boldface = retained data, Gray = potentially lost LSBs).

Digital Object Identifier 10.1109/MSP.2011.942521
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE

smallest signals may not be representable as they round to zero. This effectively decreases the dynamic range of the data set. For a compressed two's complement number, the number of sign bits is reduced and room is made to retain the bits that would have been rounded off as shown for an ideal case in Figure 1(c).

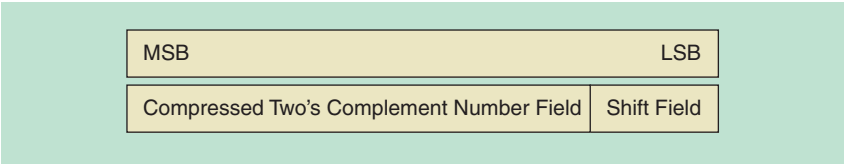
THE COMPRESSED TWO'S COMPLEMENT FORMAT

A compressed two's complement number has a predefined CF, and two data fields as shown in Figure 2 (CF is assumed).

Each member of this family of formats is identified with a different CF (one, two, three, etc.). The CF is assumed but not included in the individual bits of the data format. The CF determines how many bits each leading sign bit is to be expanded to for mathematical calculations. The shift field identifies how many digits to left shift the expanded number by. The compressed two's complement number is just a traditional two's complement number where each leading sign bit represents more than one leading sign bit when the number is expanded. The various widths of each of these fields constitute the different families in this class of data formats. The CF of a particular format generally determines the width of the shift field. The remaining bits constitute the compressed two's complement number field.

A CF of one yields a shift field of 0 b in width, and a standard two's complement number field of N b and results in a traditional fixed-point two's complement data format. A CF of two results in a shift field width of 1 b and a compressed two's complement number field of N-1 b. The minimum size of the shift field in bits is equal to the base two logarithm of the CF rounded up to the nearest integer. A CF of four results in a number field of N-2 b and a shift field of 2 b, and so forth.

An example may help. Let's illustrate a format with a CF of two and word length of five. The shift field for such a format is equal to $\log_2(\mathrm{CF})$, or 1 b in width. There are 4 b remaining for the compressed two's complement number field. Each leading sign bit is doubled, and if the shift bit is set to one, then the number is left shifted after being expanded. The resulting values are shown in Table 1.



[FIG2] **Format of a compressed two's complement number.**

With a CF of two, the 4 b two's complement number field is expanded into 8 b. This expansion obviously provides more dynamic range for the data format. What is not obvious at first glance is that a compressed data format also provides better numeric performance than a 5 b standard two's complement format would provide. Evidence of performance improvement is presented later in this article.

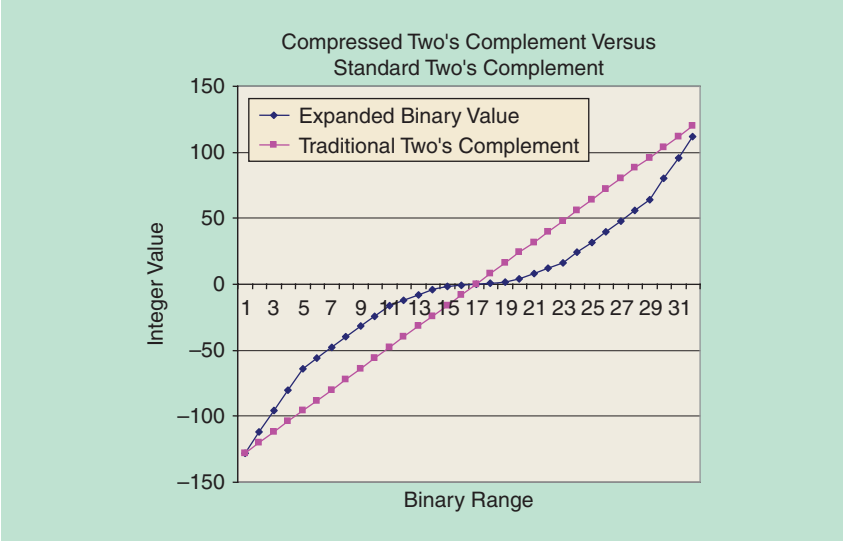
STEPS FOR COMPRESSION AND DECOMPRESSION

The steps followed for decompression with a CF of two are as follows:

[TABLE 1] A 5-b COMPRESSED TWO'S COMPLEMENT NUMBER WITH A COMPRESSION FACTOR OF TWO.

COMPRESSED BINARY VALUE (TWO'S COMPLEMENT FIELD—SHIFT FIELD)	EXPANDED BINARY VALUE	NUMERIC VALUE
0000 0	00000000	0
0000 1	00000001	1*
0001 0	00000010	2
0001 1	00000100	4
0010 0	00001000	8
0010 1	00010000	16
0011 0	00001100	12
0011 1	00011000	24
0100 0	00100000	32
0100 1	01000000	64
0101 0	00101000	40
0101 1	01010000	80
0110 0	00110000	48
0110 1	01100000	96
0111 0	00111000	56
0111 1	01110000	112
1000 0	11000000	−64
1000 1	10000000	−128
1001 0	11001000	−56
1001 1	10010000	−112
1010 0	11010000	−48
1010 1	10100000	−96
1011 0	11011000	−40
1011 1	10110000	−80
1100 0	11110000	−16
1100 1	11100000	−32
1101 0	11110100	−12
1101 1	11101000	−24
1110 0	11111100	−4
1110 1	11111000	−8
1111 0	11111111	−1
1111 1	11111110	−2

*0s are shifted into the LSB for all cases but this one.



[FIG3] Compressed two's complement versus standard two's complement for a 5-b number with a compression factor of two.

1) Double the number of leading sign bits.
2) Left justify the number into a field twice as wide as the compressed two's complement number field.
3) Pad the empty bits to the right with zeros.
4) If the shift field equals one and the compressed number is not equal to zero, then shift the number left by one bit, shifting a zero into the LSB.
5) If the compressed number is equal to zero and the shift bit is set to one, then shift a one into the LSB instead of a zero.

Similar steps are performed for formats with CFs other than two. For example, the first step for decompression of a number with a CF of four is to quadruple the number of leading sign bits. The two bit shift field of a CF = 4 format allows left shifts in Step 4 from zero to three.

The steps followed to compress a two's complement number with a CF of two are as follows:

1) Reduce the number of leading sign bits by a factor of two, rounding up for odd numbers (e.g., five leading sign bits is rounded up to three).
2) If the number of original leading sign bits is odd, then set the shift bit to one.
3) Round the resulting number to the word width of the compressed two's complement number field. (e.g., if one is compressing from 8 b to 4 b, round the result from Step 1 to 4 bits). The technique used for rounding is very important, but that subject is adequately treated elsewhere [1].

WHEN COMPRESSED TWO'S COMPLEMENT NUMBER FORMATS ARE USED ON LARGER WORD WIDTHS, TRULY IMPRESSIVE THINGS BEGIN TO HAPPEN.

4) If during the rounding process, the leading significant digit (i.e., nonsign bit) overflows into a new compressed sign bit, then the leading compressed sign bits and shift bit must be adjusted.

The data from Table 1 is plotted in Figure 3. This figure illustrates that greater precision is provided for the smallest numbers in a compressed two's complement data format. This is similar to what is accomplished with 8-b μ-law or A-law companding, and with floating-point data formatting. The main advantage compressed two's complement has over floating point is that the arithmetic is easier to accomplish in hardware, and that more precision is provided for the largest numbers in the format. This last point is not important for many problems, but is very important for DSP problems.

It turns out that for problems that require even distribution of precision, floating-point formats outperform compressed two's complement formats. DSP applications typically do not fall into this category because data is often represented as a fraction, and AGC techniques are often used to fit a problem's dynamic range into a numeric format's dynamic range. This arrangement typically gives fixed-point data formats an advantage in numeric precision over floating-point formats of equivalent word width. However, in this situation, compressed two's complement formats have an advantage over both fixed- and floating-point data formats in terms of numeric precision.

Another example will serve to hammer down this process. For a 5-b number with a CF of four, the shift field would be 2-b wide ($\log_2(4)$) and the compressed number field would be 3-b wide. During uncompression, each leading sign bit would be quadrupled, and right shifts of zero to three would place the remaining bits in their appropriate position.

The examples we have presented illustrate the process and use of compressed two's complement data formats. But we have restricted ourselves to using small word widths to improve the ease of illustration. When compressed two's complement number formats are used on larger word widths, truly impressive things begin to happen. For example, using a compressed two's complement data format with a CF of two and a 16-b word width, the dynamic range is doubled to approximately 180 dB and the round-off noise is reduced when compared with traditional two's complement arithmetic. A CF of three provides almost a tripling of dynamic range in bits (272 dB) together with improved round-off noise performance. This is illustrated in Figure 4.

IMPLEMENTATION DETAILS

It may appear at first glance that we have thrown out one of the major advantages of the two's complement

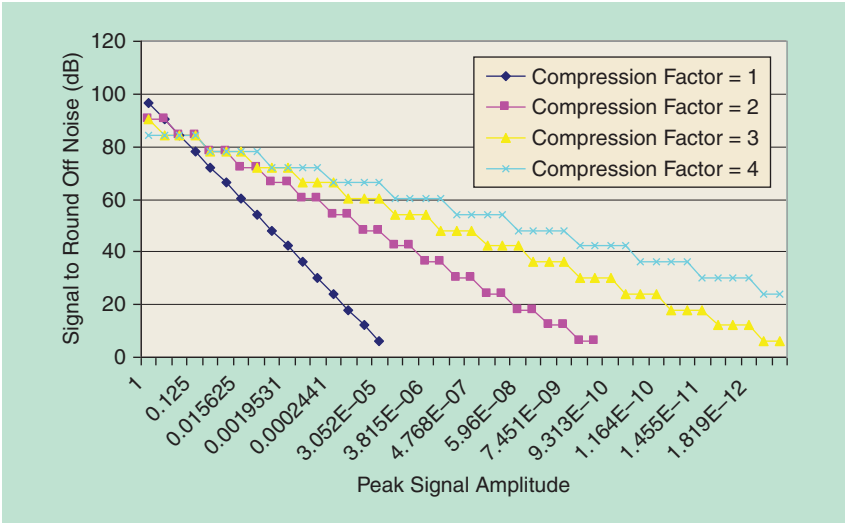
numbering systems by using a compressed format, that being the simplicity of the format. To some extent, this is true. However, we will endeavor to show that the sacrificed simplicity is not as great as first suspected. Furthermore, for many applications, the sacrificed simplicity in arithmetic implementation is more than made up for by increased simplicity in fixed-point signal processing algorithms.

Compressed two's complement algorithms can be implemented in either software or hardware. Software implementation is appropriate for low-speed applications and where data storage is emphasized over throughput. Examples of potential uses would be for stored audio on a compact disc or compressed speech. A software implementation of a compressed two's complement format with a CF of two is presented elsewhere in the form of a C++ class [3].

An appropriate hardware implementation would be within the arithmetic unit of a programmable digital signal processor integrated circuit. Such a circuit typically contains a multiply-accumulator circuit [5] for implementing digital filters. A decompression circuit should precede the traditional fixed-point arithmetic circuit where calculations are performed. When calculation is completed, data should flow through a compression circuit before being stored back into memory. The accumulator component of the arithmetic circuit would be larger than a traditional accumulator by the CF. For example, a 16-b data format with a CF of four would require a 64-b arithmetic logic unit (ALU). However, the multiplier for such an arithmetic unit would be smaller as only 14 b of precision need be multiplied. Such an arithmetic unit for a CF of 4 is shown in Figure 5.

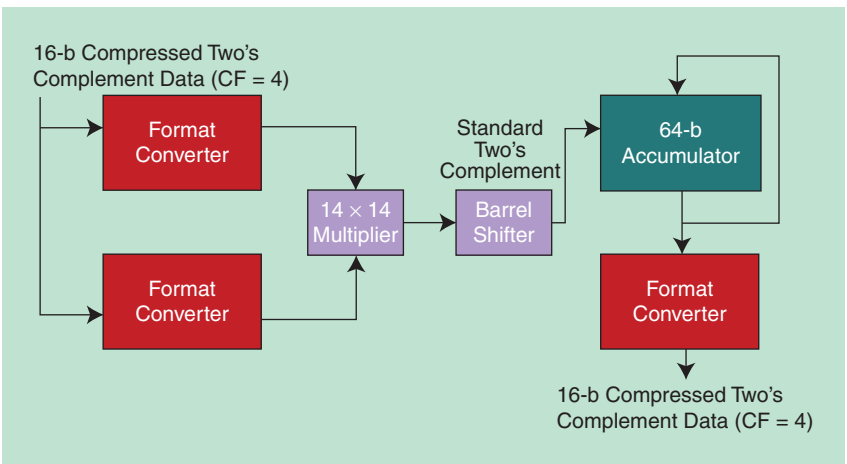
PERFORMANCE IMPROVEMENTS

Much of the information presented here to validate performance improvements has been previously published [4], but is provided here to correlate it with compressed two's complement formatting. The improvement in dynamic range by using compressed two's complement is easily calculated. In decibels, that is 3.06* (CF – 1) * [word width – $\log_2$ (CF)]. The result is a dramatic yet obvious improvement. What is less obvious is that while compressed two's complement improves dynamic range, it also improves round-off noise performance.

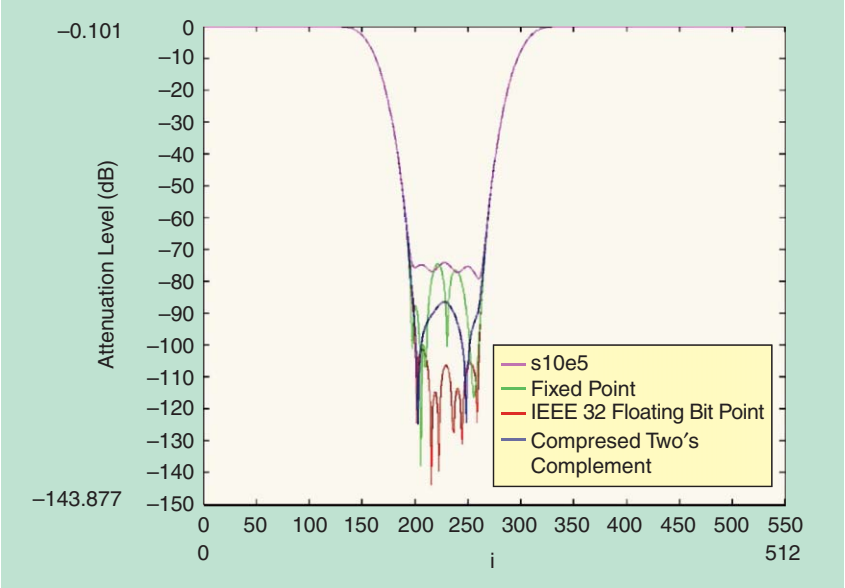


[FIG4] **Peak signal versus peak round-off noise for several small compression factors in a fractional compressed two's complement format.**

Figure 6 provides a powerful illustration of the performance gains that can be achieved through the use of compressed two's complement data formats. In this figure, the frequency spectrum of a digital notch filter is shown using four different formats. The filter was created using IEEE 754 32-b floating-point coefficients. The IEEE 754 floating-point format is shown in red. The coefficients were then converted to three other 16-b data formats and then converted back to IEEE floating point. The frequency spectrum was then plotted for the other three data formats. Traditional 16-b two's complement coefficients are plotted in green. The magenta line shows the result using a 16-b floating-point format (similar to IEEE Binary16) with 5 b of exponent and 10 b of mantissa. The purple line shows the same coefficients using a 16-b compressed two's complement data format with a CF of two. As can be seen, the rejected stop band is ten decibels lower with the compressed two's complement format than for either of the uncompressed formats. Other than conversion to and from a 16-b data format,



[FIG5] **A multiply-accumulate circuit for a compression factor of four.**



[FIG6] The frequency spectrum of a notch filter comparing the performance of various data formats.

no arithmetic was performed in the example illustrated in Figure 6.

A large DSP simulation has also been performed to validate improved performance with the compressed two's complement data formats [3]. An amplitude modulation (AM) receiver simulation was used to compare the noise performance for various data formats. The AM format was selected because this problem is well understood and still used even if in decline. This simulation contained several typical DSP algorithms that include the following: quantization to simulate 16-b analog-to-digital conversion, finite impulse response and infinite impulse response filters; demodulation; AGC; Hanning window; fast Fourier transform; and signal-to-noise ratio measurement via Parseval's theorem. AGC techniques were used following several stages to improve the native performance of the fractional format.

The simulation included the IEEE 32-b floating-point format (as a comparison baseline), together with four 16-b fixed-point formats. These were the s10e5 16-b floating point, a 16-b logarithmic format, a 16-b two's complement fractional fixed point, and a compressed two's complement fractional format with a CF of two. The simulation was performed for both weak signal and strong signal cases and both with and without the use of a single large post-multiply accumulator. Noise was not added to the simulation, so the resulting noise is a consequence of round-off errors during calculation, quantization to simulate A/D conversion, and out of band filter rejection (just over 50 dB). The simulation results are shown in Table 2.

[TABLE 2] SUMMARY OF SIMULATION RESULTS (IN dB).

FORMAT	WEAK SIGNAL SNR WITHOUT ACCUM	WEAK SIGNAL SNR WITH ACCUM	STRONG SIGNAL SNR	DYNAMIC RANGE
IEEE 754 32-B FLOATING POINT	31.97	31.97	50.53	1530
16-B s10e5 FLOATING POINT	8.44	7.90	42.06	***252***
16-B LOGARITHMIC	8.23	8.32	38.61	**385**
16-B FIXED-POINT FRACTIONAL	***13.30***	***24.93***	***44.42***	96
16-B COMPRESSED TWO'S COMPLEMENT FRACTIONAL (CF = 2)	**21.91**	**27.16**	***50.13***	181

Note: Among the 16-b formats, boldface and underlines indicates the best performer and italicized boldface the second best performer.

As can be seen from Table 2, the compressed two's complement format significantly outperformed the other 16-b formats in terms of noise performance and approached the performance of 32-b floating point for this simulation. It also provides almost twice the dynamic range of traditional fixed point.

CONCLUDING REMARKS

In this article, we have introduced and analyzed a new family of compressed fixed-point data formats for signal processing applications. These sign-bit compressed two's complement data formats are shown to provide greater dynamic range and improved noise performance over traditional fixed-point and floating-point data formats. Of course, the most desirable implementation for compressed two's complement would be in the ALU of a high-speed programmable DSP. Our results indicate that such a DSP should outperform a traditional DSP of equivalent data width in terms of algorithm performance.

AUTHORS

Manuel Richey (manuel.richey@honeywell.com) is a principal engineer at Honeywell International in Kansas where he has worked for over 25 years. He is also a computer science instructor at Fort Scott Community College and holds ten U.S. patents.

Hossein Saiedian (saiedian@eecs.ku.edu) is a professor and an associate chair in the Department of Electrical Engineering and Computer Science and a member of the Information and Telecommunication Center Lab at the University of Kansas.

REFERENCES

[1] C. Maxfield and A. Brown, *How Computers Do Math*. Hoboken, NJ: Wiley, 2005.

[2] B. Parhami, *Computer Arithmetic: Algorithms and Hardware Designs*. New York: Oxford Univ. Press, 1999.

[3] M. Richey, "The application of irregular data formats for improved performance in 16-bit digital signal processing systems," M.S. thesis, Dept. Elect. Eng. Comp. Sci., Univ. Kansas, 2006.

[4] M. F. Richey and H. Saiedian, "A new class of floating-point data formats with applications to 16-bit digital-signal processing systems," *IEEE Commun.*, vol. 47, no. 7, pp. 94–101, 2009.

[5] E. J. Tan and W. B. Heinzelman, "DSP architectures: Past, present and future," *ACM SIGARCH*, vol. 31, no. 3, pp. 6–19, 2003.

[SP]

[standards in a NUTSHELL]

Panos Kudumakis, Xin Wang,
Sergio Matone, and Mark Sandler

MPEG-M: Multimedia Service Platform Technologies

MPEG-M is a suite of ISO/IEC standards (ISO/IEC 23006) that has been developed under the auspices of Moving Picture Experts Group (MPEG). MPEG-M, also known as Multimedia Service Platform Technologies (MSPT), facilitates a collection of middleware application programming interfaces (APIs) and elementary services (ESs) as well as service aggregation so that service providers (SPs) can offer users a plethora of innovative services by extending current Internet Protocol television (IPTV) technology toward the seamless integration of personal content creation and distribution, e-commerce, social networks, and Internet distribution of digital media.

MOTIVATION

With the deployment of broadband networks enabling new ways to deliver and exchange multimedia services and the improvement of hardware performance allowing many service aspects to be implemented as Web-service software, businesses related to media services are facing significant changes. These changes are opening new business opportunities for multimedia services, such as those generated by the recent introduction of IPTV services for which several standards have been or are being developed. Examples of already developed standards are: ITU-T Q.13/16, Open IPTV Forum, Alliance for Telecommunications Industry Solutions IPTV Interoperability Forum, Digital Video Broadcasting IPTV, Hybrid Broadcast Broadband TV, and YouView.

Digital Object Identifier 10.1109/MSP.2011.942296
Date of publication: 1 November 2011

However, most of the current IPTV efforts stem from rather conventional value chain structures thus standing in stark contrast with the buoyant Web environment where new initiatives—sometimes assembling millions of users in a fortnight—pop up almost daily with exciting new features, such as Apple's and Google's APIs, enabling third parties to provide applications and services.

At the same time, we are witnessing cases where the closed delivery and content bundles offered by some operators are being either abandoned (e.g., mobile phone brands linked to a particular content service) or complemented with the possibility offered to users to freely access services (e.g., broadband, mobile, and IPTV) of their choice. The latter becomes more eminent by the appearance of new operators offering service components (e.g., cloud services) and the need for these to be interoperable.

MPEG HAS BEEN THE PROVIDER OF SOME ENABLING TECHNOLOGIES AND HAS DEVELOPED A LARGE PORTFOLIO OF STANDARDS THAT CAN BE ASSEMBLED TO PROVIDE MULTIMEDIA SERVICES.

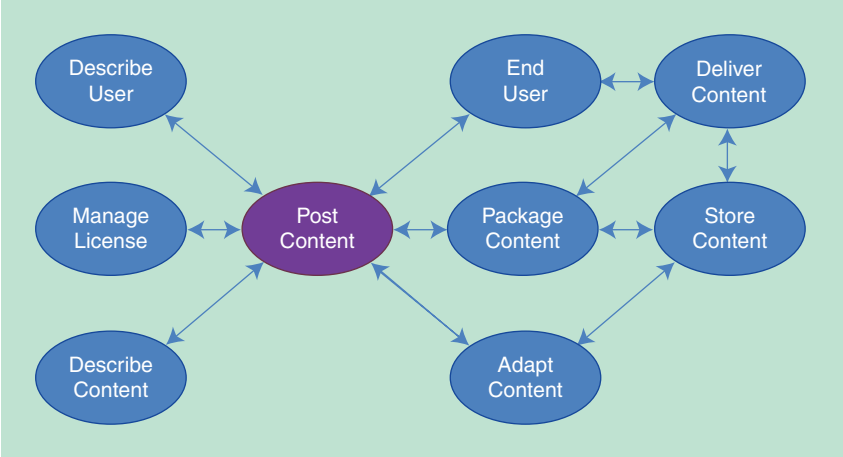
MPEG has been the provider of some enabling technologies and has developed a large portfolio of standards that can be assembled to provide multimedia services (see "MPEG technologies" under the section "Resources"). Continuing its approach of providing standards for the next generation of products, services, and applications MPEG has developed MPEG-M, a standard for advanced IPTV services. MPEG-M is based on a flexible architecture capable of accommodating and extending in an interoperable fashion many of the features that are being deployed on the Web for delivering and consuming multimedia content (e.g., Hulu, Netflix, or Apple TV), next to those enabled by the recent standard MPEG technologies (e.g., high-efficiency video coding and dynamic adaptive streaming over HTTP).

Thanks to the MPEG-M suite of standards aimed at facilitating the creation and provisioning of vastly enhanced IPTV services, it is envisaged that a thriving digital media economy can be established, where

- *developers* can offer MPEG-M service components to the professional market because a market will be enabled by the standard MPEG-M component service API
- *manufacturers* can offer MPEG-M devices to the global consumer market because of the global reach of MPEG-M services
- *SPs* can set up and launch new attractive MPEG-M services because of the ease to design and implement innovative MPEG-M value chains
- *users* can seamlessly create, offer, search, access, pay/cash, and consume MPEG-M services.

The MPEG-M suite of standards extends the devices capabilities with advanced features such as content generation, processing, and distribution by a large number of users; easy creation of new services by combining service components of their choice; global, seamless and transparent use of services regardless of geolocation, service provider, network provider, device manufacturer,

1053-5888/11/$26.00©2011IEEE



[FIG1] **A possible chain of services centered around postcontent SP.**

and provider of payment and cashing services; diversity of user experience through easy download and installation of applications produced by a global community of developers since all applications share the same middleware APIs; and innovative business models because of the ease to design and implement media-handling value chains whose devices interoperate because they are all based on the same set of technologies, especially MPEG technologies.

OBJECTIVES

The scope of the MPEG-M is to support the SPs' drive to deploy innovative multimedia services by identifying a set of ESs and defining the corresponding set of protocols and APIs to enable any user in an MPEG-M value chain to access those services in an interoperable fashion. Note that an MPEG-M value chain is a collection of users, including creators, end users, and SPs that conform to the MPEG-M standard.

Assuming that in an MPEG-M value chain there is a SP for each ES, a user may ask the postcontent SP to get a sequence of songs satisfying certain content and user descriptions (metadata). The "mood" of a group of friends could be a type of user description.

With reference to Figure 1, the end user would contact the post content SP who would get appropriate information from both the describe content SP and the describe user SP to prepare the sequence of songs according to the "mood" of the friends by using, for example, a semantic music playlist generator (see "SoundBite" under the section "Resources"). The end user would then get the necessary licenses from the manage license SP. The sequence of songs would then be handed over to the package content SP, possibly in the form of an "MPEG-21 Digital Item," the latter being a container for resources, metadata, rights, and their interrelationships. The package content SP will get the resources from the store content SP and hand over the packaged content to the deliver content SP who will stream the packaged content to the end user.

In many real-world MPEG-M value chains, SPs would not be able to exploit the potential of the standard if they were confined to only offer ESs. Therefore SPs will typically offer bundles of ESs, known as aggregated services (ASs). In general, as shown in Figure 2, there will be a plurality of SPs offering the same or partially overlapping ASs, for example, a SP offering user description services, may offer content description services as well.

Starting from MPEG-M ESs, the aggregation of services can put together a certain amount of services generating a complex MPEG-M value network, having different topologies and associating services in several ways. For example, the payment and cashing and rights negotiation ESs are aggregated to create AS#4, while content delivery and license provision ESs are both shared between AS#6 and AS#7.

ISSUING BODY, STRUCTURE OF THE STANDARD, AND SCHEDULE

MPEG-M (ISO/IEC 23006) is a suite of standards that has been developed under the auspices of MPEG.

ISO/IEC 23006 is referred as MPEG Extensible Middleware (MXM) in its first edition, and it specifies an architecture (Part 1), an API (Part 2), a reference software (Part 3), and a set of protocols to which MXM devices had to adhere (Part 4).

MPEG-M (ISO/IEC 23006) is referred to as multimedia service platform technologies (MSPT) in its second edition, and it conserves the architecture and design philosophy of the first edition, but stressing the Service Oriented Architecture character. It specifies also how to combine ESs into aggregated services (Part 5) and usage guidelines (Part 6).

The latter is subdivided into six parts:

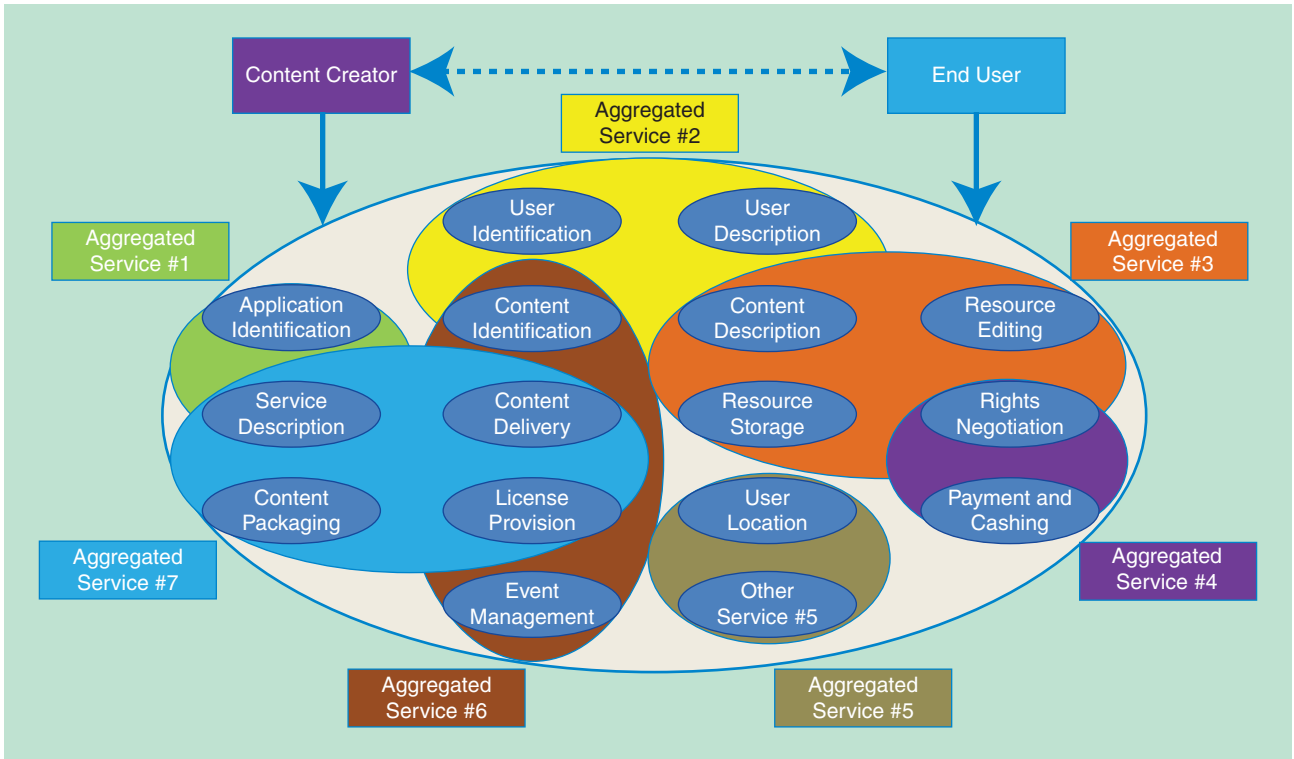
- *Part 1—Architecture*: specifies the architecture that is part of an MPEG-M implementation
- *Part 2—MXM API*: specifies the middleware APIs
- *Part 3—Conformance and Reference Software*: specifies conformance tests and the software implementation of the standard
- *Part 4—ESs*: specifies ES protocols between MPEG-M applications
- *Part 5—Service Aggregation*: specifies mechanisms enabling the combination of ESs and other services to build aggregated services
- *Part 6—Usage Guidelines*: specifies examples on elementary and aggregated services.

The specification of the MPEG-M suite of standards reached Study of Draft International Standard (SoDIS) status in July 2011, while its reference software and conformance tests are planned to be finalized in April 2012.

TECHNOLOGY

ARCHITECTURE AND SUPPORTED COMPONENTS

Next we describe the six parts of the MPEG-M suite of standards.



[FIG2] **MPEG-M standard-enabled digital media services ecosystem underpinning and supporting the activities of content creators and consumers.**

PART 1—ARCHITECTURE (23006-1)

A general architecture of an MPEG-M device is given in Figure 3, where MPEG-M applications running on an MPEG-M device could call the technology engines (TEs) in the middleware, via an application-middleware API, to access local functionality modules, and the protocol engines (PEs) to communicate with applications running on other devices by executing elementary or aggregated service protocols among them. The role of the orchestrator engine is to set up a more complex chain of TEs and PEs.

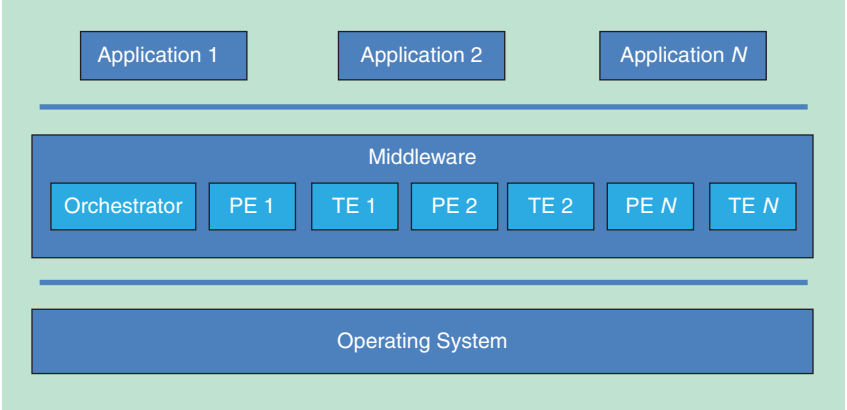
Typical TEs include those implementing MPEG technologies such as audio, video, 3-D graphics, sensory data, file format, streaming, metadata, search, rendering, adaptation, rights management and media value chain ontologies (see "MPEG Technologies" in the section "Resources").

Typical PEs include those implementing the ESs, as described earlier in the music "mood" example, such as describe user, describe content, manage license, package content, and deliver content.

The elements of the MPEG-M architecture are listed as follows:

- *MPEG-M engines* are collections of specific technologies that can be meaningfully bundled together.
- *The MPEG-M engine APIs* can be used to access MPEG-M engine functionality.
- *The MPEG-M orchestrator engine* is a special MPEG-M engine capable of creating chains of MPEG-M engines to execute a high-level application call such as "Photo Slide Show."
- *The MPEG-M orchestrator engine API* can be used to access the MPEG-M orchestrator engine.
- *An MPEG-M device* is a device equipped with MPEG-M engines.
- *An MPEG-M application* is an application that runs on an MPEG-M device and makes calls to the MPEG-M engine APIs and the MPEG-M orchestrator engine API.

In general, an MPEG-M device can have several MPEG-M applications running on it, e.g., a music or video



[FIG3] **MPEG-M device architecture. The middleware is populated by TEs and PEs.**

player as well as a content creator combining audio-visual resources with metadata and rights information. Some applications may be resident (i.e., loaded by the MPEG-M manufacturer) while some may be temporary (i.e., downloaded for a specific purpose).

When an MPEG-M application is executed, there may be "low-level" calls directly to some MPEG-M engines through their APIs and "high-level" calls such as the "Photo Slide Show," which will be handled by the orchestrator engine. The MPEG-M orchestrator is capable of setting up a chain of MPEG-M engines for handling complex operations, orchestrating the intervention and send/receive data to/from the particular chain of engines that a given high-level call will trigger, thus relieving MPEG-M applications from the logic of handling them.

PART 2—APPLICATION PROGRAMMING INTERFACE (23006-2)

This part of ISO/IEC 23006 specifies a set of APIs so that MPEG-M applications executing on an MPEG-M device can access the standard multimedia technologies contained in its middleware as MPEG-M TEs, as specified by Part 1 of ISO/IEC 23006.

The middleware APIs belong to two classes:

- The MPEG-M engine APIs, i.e., the collection of the individual MPEG-M engine APIs providing access to a single MPEG technology (e.g., video coding) or to a group of MPEG technologies where this is convenient
- The MPEG-M orchestrator API, i.e., the API of the special MPEG-M engine that is capable of creating chains of MPEG-M engines to execute a high-level application call such as "Photo Slide Show," as opposed to the typically low-level MPEG-M engine API calls.

The MPEG-M engine APIs are divided in four categories: creation APIs, editing APIs, access APIs, and engine-specific APIs.

1) *Creation APIs* are used to create data structures, files, and elementary streams conforming to the respective standards.

2) *Editing APIs* are used to modify an existing data structure, file, elementary stream to obtain a derived object still conforming to the respective standard.

3) *Access APIs* are used to parse data structures, files, and decode elementary streams to retrieve the information contained within.

WIM.TV IS CONSTANTLY ENRICHED WITH MORE FEATURES TO ENABLE MORE BUSINESS ROLES AND MAKE THE ECOSYSTEM AS LIVELY AND EFFICIENT AS POSSIBLE.

4) *Engine-specific APIs* are those that do not fall into the above categories, such as APIs for license authorization and content rendering.

[TABLE 1] ESs CLASSIFIED BY OPERATIONS AND ENTITIES.

	CONTENT	CONTRACT	DEVICE	EVENT	LICENSE	SERVICE	USER
AUTHENTICATE	X	X					
CHECK WITH		X					
CREATE	X	X			X		
DELIVER	X	X					
DESCRIBE	X		X			X	X
IDENTIFY	X	X	X				
NEGOTIATE		X			X		
PACKAGE	X						
PRESENT		X					
PROCESS	X				X		
REQUEST	X	X	X	X	X		
REVOKE		X			X		
SEARCH	X	X	X		X	X	X
STORE	X	X		X	X		
TRANSACT	X				X		
VERIFY		X	X		X		

PART 3—CONFORMANCE AND REFERENCE SOFTWARE (23006-3)

This part of ISO/IEC 23006 specifies conformance tests and the software implementation of the standard.

PART 4—ELEMENTARY SERVICES (23006-4)

This part of ISO/IEC 23006 specifies a set of ESs and protocols enabling distributed applications to exchange information related to content items and their parts, including rights and protection information.

In particular, each ES corresponds to an operation and a type of entity on which the operation is performed. Table 1 shows the ESs defined in this part of MPEG-M with the rows indicating the operations and the columns indicating the entities.

ESs can be combined in well-defined sequences to build aggregated services, both of them are called, in general, multimedia services. The multimedia services are provided by and consumed by multimedia devices in an MPEG-M ecosystem, an example of which is the advanced IPTV terminal.

PART 5—SERVICE AGGREGATION (23006-5)

The Business Process Model and Notation (BPMN) was useful toward the implementation of the MPEG-M vision, that is, the creation of aggregated services from a number of predefined ESs. The primary goal of the BPMN is to provide a standard notation that is readily understandable by all business stakeholders. These business stakeholders include the business analysts who create and refine the processes, the technical developers responsible for implementing the processes, and the business managers who monitor and manage the processes. Consequently, BPMN is intended to serve as common language to bridge the communication gap that frequently occurs between business process design and implementation.

BPMN specifies both a graphical notation and an XML representation for processes; it is a formalism to describe service workflows. Using this notation, it is possible to represent temporal events, associations, and precedences that are

combining ESs into an aggregated service. From this point of view, service aggregation is seen as an instance of a process described using BPMN.

A service aggregation can express a process flow realizing a specific task as well as a new service. The SP can expose elementary or aggregated services, constructed from several other services. Since service aggregation is a key point of MPEG-M, BPMN has been adopted because it allows efficient description of service interactions. Moreover many different aggregation topologies (not only a serial version of aggregation) and contacts among services can be quite easily illustrated employing the BPMN graphical notation.

PART 6—USAGE GUIDELINES ON ELEMENTARY AND AGGREGATED SERVICES (23006-6)

Part 6 will provide usage guidelines for a number of aggregated services. Services under consideration, for inclusion in this part of the MPEG-M standard, include selling music on the Web; buying advertisement space on video streaming services; user-centric TV programming with advertisements; professional content mash-up; IPTV content marketplace; advanced music services; personal casting of traffic events; distance learning; video-on-demand services via speech interface; and management of content adaptation.

FURTHER TECHNICAL DEVELOPMENTS

The further technical developments related to MPEG-M suite of standards is the provision of the reference software and conformance tests, as is the policy for every MPEG standard. Toward this goal a special project, Open Connected TV (OCTV), has been set up by the Digital Media Project.

The expected outcome of OCTV will not be a complete product or a running service, but a commercial-grade implementation of software—instead of the usual provided reference software—that may be used direct by implementers in commercial products and services.

Furthermore, Web/Internet/Mobile TV (WIM.TV) is an early implementation of the MPEG-M suite of standards. WIM.TV is a digital media ecosystem for diffuse trading of video content. Currently the following business roles are supported: creators, advertisers, syndicators, agents, Web TVs, end users, event minitowers, and Web banks. However, WIM.TV is constantly enriched with more features to enable more business roles and make the ecosystem as lively and efficient as possible.

RESOURCES

- Apple. (2011). iOS. [Online]. Available: http://developer.apple.com/
- Google. (2011). Android. [Online]. Available: http://developer.android.com/
- L. Chiariglione. (2011). MPEG technologies. [Online]. Available: http://mpeg.chiariglione.org/technologies.php
- Vision, Applications and Requirements for High Efficiency Video Coding (HEVC). (2011, Jan.). ISO/IEC JTC1/SC29/WG11/N11872. [Online]. Available: http://phenix.it-sudparis.eu/mpeg/
- Text of ISO/IEC DIS 23001-6 Dynamic Adaptive Streaming over HTTP (DASH). (2011, Jan.). ISO/IEC JTC1/SC29/WG11/N11749. [Online]. Available: http://phenix.it-sudparis.eu/mpeg/
- Centre for Digital Music, Queen Mary University of London. SoundBite: Semantic music playlist generator. [Online]. Available: http://isophonics.net/content/soundbite
- Information technology—Multimedia framework (MPEG-21)—Part 2: Digital item declaration. ISO/IEC 21000-2:2005. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41112
- Object Management Group. (2011, Jan.). Business process model and notation (BPMN). [Online]. Available: http://www.omg.org/spec/BPMN/2.0/
- Call for participation in the Open Connected TV project. (2011, Jan.). Digital Media Project, Doc. No. 1306. [Online]. Available: http://open.dmpf.org/dmp1306.doc

STANDARD

ISO/IEC JTC 1/ SC 29/ WG 11 Information Technology—Multimedia Service Platform Technologies (23006):

- Part 1—Architecture (23006-1), N11934, Mar. 2011
- Part 2—MXM API (23006-2), N11936, Mar. 2011
- Part 3—Conformance and Reference Software (23006-3), N12148, Aug. 2011
- Part 4—ESs (23006-4), N12149, July 2011.
- Part 5—Service Aggregation (23006-5), N12150, Aug. 2011
- Part 6—Usage Guidelines on Elementary and Aggregated Services (23006-6), N12151, July 2011.

SERVICE

WIM.TV. (2011). [Online]. Available: http://www.wim.tv/wimtv-webapp/wimTvHome.do

AUTHORS

Panos Kudumakis (panos.kudumakis@eecs.qmul.ac.uk) is a research manager at qMedia, Queen Mary University of London, United Kingdom. He is coeditor of ISO/IEC Multimedia Service Platform Technologies—Part 1: Architecture (23006-1) and has been an active member of the ISO/IEC MPEG Standard Committee since 1998.

Xin Wang (xwang@huawei.com) is the head of the Multimedia Systems Lab at Corporate Research of Huawei Technologies, Santa Clara, California, United States. He is cochair of the MPEG-M ad hoc group and coeditor of ISO/IEC Multimedia Service Platform Technologies—Part 1: Architecture (23006-1).

Sergio Matone (sergio@cedeo.net) is a digital media technologist at CEDEO.net, Torino, Italy. He is coeditor of ISO/IEC Multimedia Service Platform Technologies—Part 2: MXM API (23006-2), Part 4: ESs (23006-4) and Part 5: Service Aggregation (23006-5).

Mark Sandler (mark.sandler@eecs.qmul.ac.uk) is the head of the School of Electronic Engineering and Computer Science and director of qMedia, Queen Mary University of London, United Kingdom.

SP

[dsp FORUM]

Fatih Porikli, Al Bovik, Chris Plack, Ghassan AlRegib, Joyce Farrell, Patrick Le Callet, Quan Huynh-Thu, Sebastian Möller, and Stefan Winkler

Multimedia Quality Assessment

This *IEEE Signal Processing Magazine* forum discusses the latest advances and challenges in multimedia quality assessment. The forum members bring their expert insights into issues such as perceptual models and quality measures for future applications such as three-dimensional (3-D) videos and interactivity media. The invited forum members are Al Bovik (University of Texas), Chris Plack (University of Manchester), Ghassan AlRegib (Georgia Institute of Technology), Joyce Farrell (Stanford University), Patrick Le Callet (University de Nantes), Quan Huynh-Thu (Technicolor), Sebastian Möller (Deutsche Telekom Labs, TU Berlin), and Stefan Winkler (Advanced Digital Sciences Center). The moderator of this forum is Dr. Fatih Porikli (MERL, Cambridge).

Our readers may agree or disagree with the ideas discussed next. In either case, we invite you to share your comments with us by e-mailing fatih@merl.com or spm.columns.forums@gmail.com.

Moderator: Let's start our discussion. Here is the first question on philosophical aspects: How would you define "objective quality" and "perceptual quality?" What are the attributes that characterize the picture (signal) quality?

Quan Huynh-Thu: I think that, generally speaking, "quality" is fundamentally the result of a human (and therefore subjective) judgment based on several criteria. Some of these criteria are measurable as they can be based on intrinsic information of the signal to judge, while other criteria are the result of cognitive processes integrating information beyond just the signal that is being judged (e.g., context, experience). So in a sense, quality is always perceptual. Now, the distinction I see concerns the algorithms (or metrics) that are designed to compute a measure of quality. I usually make the following distinction: "objective quality" is the value calculated by a computational model, whatever that model is. In other words, computational models produce an objective quality as a prediction of (subjective) quality. The simplest one, which is still widely used in the literature, is the peak signal-to-noise ratio (PSNR). However, such a model is known for not representing faithfully the human visual response, e.g., masking processes. I would then use "perceptual quality" to either represent subjective quality itself (i.e., as judged by human beings) or represent the value calculated by a computational model when its prediction is closely correlated with human judgment. Now, obviously, the difficulty and grey area is in how closely this should be. I guess I don't have a simple answer to this. Experts use different kinds of statistical analyses to quantify and characterize the performance of objective quality assessment models, but there is no defined or agreed simple threshold to define whether a model computes an objective quality that correlates enough with subjective quality so to be considered as perceptual quality. This is where it becomes a bit philosophical.

As expressed previously, I think that some of the attributes that characterize the picture quality can be objectively described from the extraction of features in the signal. The particular features that need to be measured depend highly on the type of signal to judge (e.g., picture, video, voice, and audio) but also on the context or types of processing/degradations that will impact the quality. Generally speaking, subjective experiments can be conducted to identify the particular attributes that are highly correlated with the human judgment of quality in the particular context of interest. Then, based on that knowledge, algorithms integrating the computation of these features can be developed.

Sebastian Möller: I think that there is no such thing like "objective quality." In my opinion, quality is the result of a perception and a judgment process, during which the perceiving human compares the features of the perceptual event, which is commonly provoked by the physical event, i.e., the signal reaching her perception organs, to the features of some internal reference. This reference reflects the expectations, experience, but also the transient internal states like motivation, emotions, and so on. Quality is thus inherently "subjective," determined by the reference of the perceiving human. It "happens" in a particular judgment situation. Still, scientifically speaking, we need to make the perception and judgment process as "objective" as possible, thus independent of the experimenter who performs the measurement.

If we want to find out about the attributes of quality, we need to identify the underlying features of the perceptual event. When judging transmitted speech under laboratory conditions, multidimensional scaling experiments as well as semantic differential techniques suggest that there are four to seven features, including sound quality, continuity, (absence of) noisiness, and loudness, to name just a few. However, additional features might arise if one digs further into one of these dimensions, or if experiments are carried out under more

Digital Object Identifier 10.1109/MSP.2011.942341
Date of publication: 1 November 2011

1053-5888/11/$26.00©2011IEEE

realistic conditions. In this case, qualitative methods are a good starting point to identify relevant features.

Chris Plack: I disagree with this. Quality can be defined (by the Merriam Webster dictionary) as "an inherent feature" or "a property," or "a degree of excellence." Both of these can be objective. Clearly, when comparing systems, someone has to decide which measurable qualities are important to excellence, but that doesn't make the qualities themselves subjective.

If it is measurable using a physical instrument, then it is an objective quality. If measurement requires a response from a human observer, then it is a perceptual quality. For example, an image with a higher spatial resolution might be considered of higher quality, and hence system quality can be compared in terms of the objective quality of resolution. On the other hand, we might also ask participants to make a subjective rating of the "crispness" of an image, which is a perceptual quality.

Sebastian Mdler: For me, the spatial resolution would be a metric for performance. The same holds for temporal resolution, bandwidth, etc. All these are entities that can easily be measured with a physical instrument. All may also have an influence on perceived quality, but they would not be qualities in my point of view—I would reserve this word for the perceptual entity. Having said this, it is also obvious that I disagree to quality being an "inherent feature" or a "property" (even if common usage of the term suggests so)—you could say that spatial resolution is an inherent feature, but it has to be judged by a human before being linked to quality.

Quan Huynh-Thu: I would tend to agree with Sebastian concerning the idea that quality is the result of a human judgment, as I wrote in my first postanswer. One of the reasons is that the concept of signal quality involves a reference point, which is mostly subjective.

Spatial resolution, temporal resolution, and bandwidth are all objectively measurable features but not qualities per se. In the example Chris gives, a higher spatial resolution image may be judged with higher quality but that is not necessarily so. It also depends on the other attributes/features. For example, at a given bit rate, encoding an image/video at resolution R1 versus resolution R2 with R1 $>$ R2 will not necessarily produce a higher perceived quality; it all depends on the bit rate used and on the values of R1 and R2. At low bit rates, it is possible that an image/video at lower resolution will be judged as being of higher quality but, at very high bit rates, the reverse judgment of quality will likely happen.

Today, in Brittany, the weather temperature is 18 °C and rainy, although it is supposedly summer. Yesterday, it was 14 °C … so usual language could say that the weather is of better quality today but truly the only thing you one can say is that today is warmer. Yesterday was colder but sunnier, so but in my opinion the weather was of better quality yesterday because the sun has a bigger effect on me than temperature, but a Breton (who doesn't care if it rains 365 days a year and will complain that it is too hot when it's over 20°) may say or think otherwise as his/her reference is different.

Chris Plack: It depends on definition, of course. However if you define quality as a property or feature of something (which is one of the most common definitions), then clearly quality can be measured objectively.

OK, in the example I gave, if you fix the bit rate, then perceptual quality may not be predictable from resolution. However this is nit-picking in my view. In some cases, there may be an interaction between objective qualities to determine perceptual quality, but that doesn't deny the fact that there are objective qualities. Or that quality can be measured objectively. Would a better example be bit rate itself?

Patrick Le Callet: With increasing interests of the field witnessed by an explosion of literature, the definition of those concepts deserves clarification, as some meanings have been progressively lost. At this point, it is worth including also the concept of "subjective quality." Initially, "subjective" and "objective" have been introduced to differentiate two different ways to assess quality. Subjective quality assessment means measuring quality with humans, requiring experiments to get their opinions. Objective quality assessment refers to a metric that provides numbers in a repeatable manner. It covers a simple case of a signal attribute as a complex algorithm that could mimic human perception. Perceptual quality is related to the quality perceived by humans, involving sensation, perception, cognition of and context of use.

Coming back to quality assessment, by definition, subjective quality assessment should be able to capture perceptual quality. This is not necessarily the case for objective quality assessment. Getting some numbers computable in a repeatable manner is what is needed in many fields and is practically more useful. An objective quality assessment should provide this; nevertheless, the question remains open as an objective quality assessment tool is not necessarily able to reflect the perceptual quality in all contexts.

Attributes that characterize signal quality are numerous; their perceptual values are mainly context dependent. (I support Sebastian's point of view regarding this part.)

Ghassan AlRegib: As mentioned in the discussions so far, depending on the usage and the application (or context), one could define these two quality measure differently. In some cases, one might need an objective quality, which is obtained via computational models. In other cases, one needs the true perceived quality. In the former, I am thinking more in terms of applications where a "machine" needs to have some number to describe the quality of a signal. In the latter, the human wants to "judge" a signal. Of course, there has been a tremendous effort in coming up with models that generate an "objective quality" that simulates what a human brain does when judging such a signal. I tend to believe that in subjective quality, humans use their background experiences, perhaps some references, and many other factors while judging the quality of a signal. They do not tend to come up with a quality that fits a certain application. On the contrary, an "objective quality" measure or a metric has a target application (e.g.,

compression, streaming, etc.) and tries to come up with a number to be used in the target application. I think this viewpoint is in line with Sebastian's statements on "inherent property" and how they are linked to quality.

What I am saying is that subjective quality and objective quality are two different things; it is not the measure that is subjective or objective, it is the "quality" that is either subjective or objective. They of course differ in the mechanism but they also differ in the constraints, the factors, and the usage. There are some attributes in the signal and there are attributes in the application and the usage of the quality that affect the choice of objective versus subjective quality measures. This becomes clearer when we developed and tried to measure the quality in 3-D video and in human interaction applications. If you give a person a job to perform (e.g., tie a shoelace) in virtual reality using state-of-the-art interactive devices (e.g., data gloves) and assign an evaluator to watch the virtual action, how would the quality assigned by the evaluator differ from an objective quality measure? It is very complex signal and even to define a "good" shoe lace tying is a very tough task to start with. Similarly, a single monocular cue such as color or texture may have some impact on the quality of a two-dimensional (2-D) video, but they have catastrophic impact on the quality of a 3-D video. Nevertheless, humans may not notice such implications in a 3-D video, while they notice it clearly in a 2-D video. Back to the example of "high resolution means higher quality." In our lab, we ran many experiments with several subjects over a number of displays (a couple of years ago) to understand the impact of resolution, frame rate, texture content, etc. on the perceived quality of ultrahigh-definition (HD) videos as well as 3-D videos. High resolution gave "good experience" for a short period of time but ultimately the quality of the content (color, sharpness, texture etc.) decided on the perceived quality of videos. Limited resolution means, in my opinion, that we can hide several artifacts from human eyes. This takes me back to the discussions on how inherent features are linked to quality but quality is not an inherent feature or a property.

Al Bovik: Since the aim of an objective quality model is to predict, as accurately as possible, perceptual quality, then I will take the position that they are the same thing, in the sense that the goal is that they be the same thing. Naturally, we are not there yet; objective models deliver numbers, and objective experiments do the same, whereas visual quality is something ineffable, of infinite dimensions, and varying with situation, context, space, and time.

Yet we are on the right track toward solving these problems in psychology and engineering. As several others have expressed, perceptual modeling is a key ingredient to success. Low-level models of cortical and extra-cortical processing, as used in the past by Beau Watson, Stefan Winkler, and many others enable us to define perceptually meaningful attributes of visual signals. Similarly, statistical models of the naturalness of visual signals, provably regular over very large, generic, and holistic collections of signals, continue to provide a dual and fundamental basis for finding fundamental visual signal attributes useful in developing objective models that predict subjective responses.

I think the value of a forum like this is to suggest directions where work is needed: Going forward, there are recent innovations being made in developing perceptual models that have not yet been exploited in visual quality assessment model design. For example, recent work on the nature of responses in the cortical area V2, and how they relate to attention and visual resolution may prove to be quite valuable. Eero Simoncelli has been doing interesting work in this direction. Further along, deepening our understanding of recognition processes in inferior cortex would greatly accelerate our broader understanding of visual quality. If I were to pick an immediate problem whose resolution would have a tremendous impact on objective algorithm work, it would be the development of a consistent, regular model of the statistics of natural videos. We do not yet have a useful one that complements highly successful still image models. Whoever succeeds at developing such a model will deeply impact vision science as well as video engineering.

Moderator: Are the perceptual and objective quality measures sufficiently general to perform reliably over a very broad set of typical content? What makes the visual signal so complex to analyze and evaluate? Is it the human perception or the composition of the signal itself?

Sebastian Möller: As mentioned above, it is the internal reference that decides on quality. As an example, the same audio signal will be judged completely differently if you listen to it through a telephone connection or through your high-fidelity chain. The signal is the same, but the reference is different, and thus your quality judgment. Models for quality prediction struggle with this, as they commonly have only a marginal representation of the internal reference.

But subjective testing also reaches its limit here: If you perform an auditory quality test under narrowband conditions, the maximum rating for the clean narrowband channel will be around 4.5 (i.e., between excellent and good) on your five-point mean opinion score (MOS) scale. For the same experiment carried out under wideband conditions, the maximum rating for a clean wideband channel will also be around 4.5. Thus, network operators ask why they should introduce wideband when it results in the same MOS. The answer is that the clean narrowband channel would roughly obtain a 3.5 (i.e., a rating between fair and good) in the wideband test.

Quan Huynh-Thu: From having researched video quality prediction models for many years, I don't believe in an objective quality assessment model that does it all, either over a broad range of visual content and/or over a wide range of degradation types. Best-performing models will typically behave well for some conditions and not so well for others. On the other hand, models can gain a lot in accuracy and robustness (when their performance is compared to subjective judgment) when they are tuned for a

specific context or type of content. A simple example is a model initially developed to handle both coding artifacts and transmission errors but is applied in a context to measure quality of encoded content in a head end. Obviously the model was developed to handle this scenario but the fact that it was also developed to take into account block/slice distortions due to transmission may introduce more false positives/negatives when measuring just coding artifacts. So the more general model is not always the most suitable model. Another example is videoconferencing versus video streaming. These two applications both use an audio-video signal but the context and type of content are so different that a model that was developed for video streaming is unlikely to work that well on videoconferencing, unless it was redesigned or readapted for that. In particular, the way we focus on human faces (which predominantly appear in videoconferencing) makes "talking-head" or "head-and-shoulder" content very different to other types of content.

Today, we have very sophisticated tools to compute a broad range of features to characterize/analyze the signal composition. So I believe this is not really the fundamental issue even if there is definitely still room for improvement, especially as new features may become more relevant in new scenarios. However, the integration of the information in our brain and the complex cognitive mechanisms used to come up with that single judgment of quality value are the critical points that we can't yet model robustly, especially when context and experience can have such a significant impact on the final judgment.

I think that we are currently able to predict subjective quality using computational models but clearly defining the context for which this model was developed and knowing the limitations of the model are as crucial as the performance itself.

Patrick Le Callet: I see two reasons: human perception of course and the context of use (environment, expectations, personal engagement). Both cannot be disconnected and difficult to model. It is still possible to handle part of them making right assumptions and approximations. But we are far away from a universal perceptual objective quality assessment tool, and we should remain humble regarding this challenge. Nevertheless, literature demonstrates that in some well-defined contexts, we are able to define an efficient tool. That leads to the question of reliability; one often forgets a key principle: an objective quality metric can be reliable from a perceptual point of view only if it has been validated using subjective quality assessment results. Subjective quality assessment is neither trivial, in some conditions, it might require some research efforts to design right protocols being to capture the right perceptual quality values. I am often surprised to see how some Communities that urgently need objective quality metric are keen to use new metrics violating the key principle for their proper needs. The quality assessment community should probably deliver a more cautious message regarding the good usage of objective quality metric mentioning its reliable context.

Al Bovik: Having spent most of the month of July in the galleries of Paris, Florence, and Rome, I continue to be impressed by the thought that visual aesthetics is the final ingredient of visual quality. Further, that aesthetics rely on what others here have referred to as content and context; viz., the Sunday couch football addict's aesthetic judgment is directed towards how well he or she can feel immersed in what he or she is watching, without distractions of distortions or poor camera work. How we capture the aesthetics of visual signals will be one of the great challenges of this field further out.

With all that said, and shifting my thoughts leftward, certainly some objective algorithms perform well over broad classes of content (especially full reference algorithms) but I agree with others that the experts in each domain—medical, military, mobile, and so on—will need to make these judgments.

Moderator: How successfully do the quality evaluation models emulate the human (visual, audio) perception? What are their limitations? How would you design a model if you had all the computational (and human) resources?

Sebastian Möller: In the speech quality domain, researchers have tried to model many of the known processes that are relevant for monaural auditory perception into quality prediction models. Disappointingly, some standard models [such as the Perceptual Evaluation of Speech Quality (PESQ)] have shown that an exact modeling of human auditory perception does not necessarily lead to accurate quality predictions—the models sometimes performed better if they deliberately violated human perception.

This shows that modeling human perception is not enough: Two other important aspects are missing. The first is the reference that we have already discussed. In a full-reference model, we usually take the undistorted (clean) signal as the reference, although we try to model an absolute category rating test, i.e., a test where the human listener does not have access to the clean signal for comparison. The human reference (reflecting long-term experience) might thus be different from the reference signal used in the model. This has led to proposals where a modified or idealized version of the reference is used for comparison, with some good results. The second point that is not yet appropriately modeled is the judgment process, in particular when it relates to an experience that is formed over a longer period of time. So-called "call-quality models" have addressed this point, but so far only for speech applications.

Joyce Farrell: Reference-based image quality metrics derived from models of human vision are typically designed to predict threshold judgments, such as the visibility of annoying artifacts. These metrics are relatively successful in predicting subjective judgments of image quality that are primarily driven by the visibility of noise, blur, blocking artifacts, color differences, flicker, and other types of distortions.

Reference-based metrics were not designed to predict suprathreshold judgments of image quality, such as one's preference for different types of image

enhancements. Moreover, these types of judgments are influenced by cultural norms and are thus more variable across diverse populations. Therefore, it should not be surprising that metrics based on human vision models will be less successful in predicting subjective preferences.

On the question of "How would you design a model if you had all the computational (and human) resources?" I am working with colleagues at Stanford University to develop a model of the human visual system that includes human optics, photo-detector gain, retinal ganglion processing, and visual cortical processing. This is an ambitious project. My hope is that we will eventually be able to design experiments that isolate these different stages in human visual processing and determine how they influence our judgments of image quality.

Quan Huynh-Thu: Some full-reference quality models proposed in the literature and included in existing standards have shown very good performance in predicting subjective quality. However, existing models are still limited to modeling the low-level processing of the human brain, i.e., they model the visibility of the artifacts and cannot yet model robustly the (cognitive) integration process that ultimately form the subjective judgment of quality. This is in my opinion the most difficult part of the modeling because the low-level part can usually be translated into signal processing techniques, but these techniques cannot easily represent the integration part. The second limitation is the context. A model usually integrates parameters that are tuned to fit subjective databases that represent subjective quality collected in a specific context. Application of a model tuned to a context to another context is therefore not straightforward and actually should be discouraged. The context relates both to the types of degradations for which the model was designed to handle but also to the type of reference signal. Changing the reference signal will usually break the model.

Stefan Winkler: A complementary question to this would be, how much do quality models actually need to emulate human vision? Of course, in terms of outputs, we want the models to come as close as possible to subjective experimental data. However, how much actual vision modeling as such is really necessary to achieve that? Quality models using a few signal parameters and fitting functions can do surprisingly well in many circumstances.

Patrick Le Callet: I'm fully supporting this point of view. Once again, this is a matter of applications or context. From the human vision modeling perspective, applications to image and video quality assessment have been mostly limited to sensation stages, e.g., very early stages of human vision that are quite accurate to predict nearly visual threshold impairments. This is very useful in some application scenario;, the Medical Imaging Perception Society community is a good example, but when distortions deal with supravisibility threshold, this is still questionable. Applying perception and cognition concepts with a bottom-up approach is so complex than it is currently more comfortable to adopt a top-down way to proceed. This comes with some assumptions that should limit the application scope but on the other hand lead to some nice performances, as long as we remain in the scope.

Nevertheless, trying to get more generic models to be able to support more application fields is a holy quest that should help us to consider human vision much beyond than sensation stages, including higher perception level and cognition.

Stefan Winkler: To me, the question is not so much how to design a model if you had all computational resources (my tongue-in-cheek answer: build 15 or more artificial brains). The much more important question is, how much better can vision-based models actually do, and do the (often incremental) improvements over traditional approaches justify the added complexity? I believe this point is essential from a practical model usage perspective, and it is also one of the reasons the acceptance of quality models by other (e.g., video coding) research communities and engineers has been disappointingly low.

Al Bovik: Regarding images and video, I think that full reference models are doing quite a good job at predicting the quality of generic signals. Naturally there is room for improvement going forward, especially in specific application domains requiring certain types and levels of performance.

But the full reference concept is terribly limiting. To me, the Holy Grail is quality assessment (QA) models that "blindly" predict visual quality without reference, and for that matter, without knowing the distortion in advance. We and others are making good progress on this for still images. For videos, and for 3-D, the problem remains elusive since our statistical signal models are undeveloped.

If we had all the resources we needed? This is easy: quality assessment is a problem in predicting behavioral psychology using video engineering tools such as sparse and efficient representations, quantitative perception models, and machine learning. But we are lacking data. Given unlimited resources, I would conduct extremely large-scale human studies of time-dependent visual response to videos of highly diverse lengths, content, distortion types, distortion durations, and other variables. In my view, one of the great challenges is understanding how quality perception changes over time, how it relies on visual memory, and how temporal variations in quality modify quality perception over wider time scales. We will soon release a nice database and human study that explores these issues, but, alas, the videos are of usual 10–15 s presentations.

Moderator: What would be the most critical granularity for visual quality: pixel, block, frame, or sequence? What would be the similar analogy in audio quality?

Quan Huynh-Thu: Yes, this depends highly on the application and mostly what people do with the quality value. Is it for detecting a severe degradation even if it is very short in time? Is it to measure average quality on an aggregated number of communication channels? Does it need to be every 10 s or every minute? Ultimately, this depends on the usage and

the level agreement between the party offering the video service and the party receiving the video service.

Stefan Winkler: As Quan already pointed out, granularity is highly application dependent. An encoder needs a very different granularity (perhaps block or even pixel-based) than the chief executive officer of a cable operator who is only interested in high-level reports (weekly or monthly). In fact, granularity is a bigger issue than people commonly realize. How do you even design a subjective test for block-level quality measurement, or for aggregating results over weeks of data, or perhaps over a number of different programs? How do short quality degradations translate into longer-term perception of quality? We know very little about how humans do these things, much less how to model it.

Patrick Le Callet: I fully agree and would like to add that is mainly due to our current lack of effort to design right subjective experiments to better understand those mechanisms. When we deal with subjective quality, usual methodologies are suitable to get short- and mid-term judgment that involves both spatial and temporal pooling. Reduced granularities are out of the game and consequently there is so far only very limited ways to construct ground truth at these levels for quality metrics.

One side additional question is: What do pixel, block, and slices mean from a human vision point of view? When designing quality metric, it is usually better to be able to translate physical parameters into perceptual space to improve robustness across various viewing conditions.

Stefan Winkler: I'd also like to raise another issue here (and my apologies for raising more questions than providing answers), on the topic of applications, or as Quan put it, what people do with the quality value.

Monitoring and comparing things is nice, but ultimately not all that exciting. If you find out something is bad, you want to know how to make it better, and by this of course I don't mean the trivial answer of "increase the bit rate" or similar. To ultimately be useful and successful, I believe quality models should be integrated in all kinds of image/video optimization and feedback loops, but very few are designed for that. Most research still focuses on the measurement aspect alone, completely disregarding the "what to do with it" issue.

This question also brings us back to granularity, not only on a temporal level, but also in terms of model output. For the type of usage I just described, having overall quality or MOS alone is insufficient, because it does not answer the question of what has gone wrong in the system, much less the question of how to fix it. Models that quantify specific artifacts go some way in addressing this, but not in a rigorous manner, in the sense that "blockiness" may be useful as another perceptual quantity, but it doesn't necessarily pinpoint a specific "culprit" either (Was it the encoder? Was it the bit rate or the content? Was it the network? Was it a loss or a delay issue?).

Joyce Farrell: I can't agree with you more Stefan. Engineers want to know how to minimize distortions and the only way they can do that is to identify the source of the distortion. As you so aptly put, they want to "fix" what has gone wrong in a system.

In the 15 years that I worked at Hewlett Packard Laboratories, I found that the people who were most vocal about the need for metrics worked in the marketing departments.

I am reminded of an anecdote someone once shared with me. Many years ago, a colleague of his was given the job to evaluate models that predict the weather. After months of research, his colleague concluded that the models they had at that time did not predict weather better than the rule "Tomorrow's weather will be like today's weather." His colleague was in the army, and when he turned in his report, he was brought before his superior, who explained to him that when commanders make a decision about a military maneuver, they cannot say that they based their decision on the theory that tomorrow's weather will be like today's weather. Luckily, in the last 50 years, we have made a lot of progress in predicting the weather. And in the future, I am hoping that our metrics will perform better than the PSNR.

But back to our discussion, I agree that engineers do not find metrics to be useful unless they help to diagnose a problem. After many years of working on image-quality evaluation in industry, I now have the time to collaborate on the design of system simulations tools (such as digital camera and display simulators) to help engineers identify the impact that different components in the image processing pipeline (from capture to processing to transmission to display) have on the final image quality. Simulation tools, in combination with metrics, can help engineers optimize the design of imaging devices.

Sebastian Möller: For speech quality, we have two projects running in Study Group 12 of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T SG12) who are dealing with a diagnostic decomposition of speech quality. One is called perceptual approaches for multidimensional analysis (P.AMD) and tries to find estimations of perceptual speech quality; the other is called technical cause analysis (P.TCA) and tries to identify the technical causes of perceptual degradations. The discussion about how many perceptual dimensions we actually need is lively, and so is the discussion about whether perceptual dimensions or technical causes are more informative—perhaps you have some ideas about this?

One of the reasons for going to perceptual dimensions instead of technical causes (at least from my point of view) was that perceptual dimensions were expected to be more stable when new transmission techniques become available. Is this the right assumption?

Al Bovik: Ultimately we should try to deploy models that resonate with perception as much as possible, but I think this can be done to a good practical approximation with any of these levels of granularity. We use them all to develop algorithms appropriate for different applications scenarios.

Ghassan AlRegib: As mentioned, this depends on applications. Humans' perception occurs at many levels so

granularity and perhaps at the same time. The scope or the goal of the quality at each granularity level is different from the other levels but collectively they serve a high-level quality that is defined and required by the application. I think the correlation among the quality measures at various granularity levels is very important and needs to receive more investigation.

Moderator: What kind of errors are more disturbing: transmission errors or compression errors? How would they be balanced?

Sebastian Möller: This is the main task of a good quality prediction model, that it is able to tradeoff between different types of impairments reflecting human perception and thus can be used for taking decisions on the optimum balance. Severe transmission impairments will normally be more disturbing than compression impairments, in case that they lead to a complete loss of information, like fading in mobile speech communication, or freezing in IPTV. In addition to transmission and compression impairments, interactive services commonly also show impairments in the source material to be transmitted, such as background noise and reverberation with speech, or bad lighting conditions with video telephony. These might be even more disturbing than compression artifacts.

Joyce Farrell: I agree with Sebastian that the value of reference-based metrics is to evaluate image quality tradeoffs to find the optimal balance of different types of image impairments. Metrics based on human vision are ideal for this task because they are designed to quantify the visibility of such impairments.

Quan Huynh-Thu: More than the level of error itself, variation of error level is really annoying. If video quality is really high but video gets corrupted regularly by slice/block errors, this becomes really annoying. Conversely, if coding quality is low but remains stable, this may not be as disturbing as a video where coding quality is high but keeps dipping down regularly, due to, say, network adaptation.

Al Bovik: In simulation, any of these can be made perceptually disturbing to any degree, but in practice uncorrected transmission errors, such as packet losses, are quite annoying. As bandwidths increase, compression and resolution-related artifacts will become less significant; but as video traffic increases and wireless video continues to take hold, transmission errors in all flavors will become increasingly problematic. Keep in mind that industry predictions are for wireless video traffic to increase 5,000% or more over the next few years.

Ghassan AlRegib: Compression errors and transmission errors will continue to exist and affect the quality of received media. New types of media (e.g., 3-D video) will always require new ways to compress the content with new models and new algorithms. Also, as the bandwidth increases, the competing streams are increasing rapidly and there will always be transmission errors, especially with the rapid increase in the number of devices consumers have to access online media.

Moderator: How would social networking, collaborative tagging (in large multimedia databases such as Flickr and Picasa) and monitoring Internet browsing of content consumers change the way we evaluate the multimedia quality?

Quan Huynh-Thu: Social networking and Internet browsing are application scenarios that encompass factors that are way beyond the concept of multimedia quality as currently addressed by the quality assessment community. In those two scenarios, the notion of context, task, and interactivity are so important that the quality of the signal itself becomes only a very marginal component of the problem of quality assessment and may not be that relevant anymore as quality of experience (QoE) is what matters rather than just multimedia quality.

Patrick Le Callet: The development of such Internet services has totally modified the rules. Today, it is almost as easy to be a content producer as a content consumer. Moreover the variety of way to produce and consume has exploded, and the context of use is following. A possible consequence could be that users will start to be more educated in terms of quality requirements. Moreover, the image quality assessment community is considering the user as a watcher only while he tends to be much more than that. The new offers on interactivity (e.g., easy browsing, annotating, editing …) are some of the key factors that have participated to the emergence of consumer-producer users. I agree with Quan that QoE is what matters, as long as it is well defined and associated to a proper service and context of use.

Stefan Winkler: User-generated content is so different from professional content that we generally consider for quality assessment that it also brings with it completely different aspects about the meaning and importance of quality. Personal meaning (your family or friends are in a picture, for example) as well as timeliness (an image from a phone camera during a trip that is shared right away may be much more valuable than a high-quality picture uploaded from home three days later) often completely outweigh any quality aspects.

Joyce Farrell: The interest in visual saliency is in part driven by our desire to know what grabs consumers' attention in an image. In this sense, monitoring Internet browsing has already changed how we evaluate multimedia quality.

We need ways to evaluate the user's experience in real time. The MOS score is limited in how it captures the ups and downs of our Internet experience. Progress in both electroencephalography and functional magnetic resonance imaging make it possible to measure the user's brain activity in real time. This may become a useful evaluation tool in the future, and it is the focus of several research projects at Stanford.

Al Bovik: This is also easy. It creates a tremendous opportunity for data gathering and model development. If we are able to gather large amounts of data on human judgments of visual quality from high-traffic Internet sites (by simply asking users to supply ratings), then even in such unconstrained environments (far from the psychometric controlled viewing conditions usually demanded), the amount of data collected should prove invaluable. I think someone coined the

term "social quality assessment." We are pursuing such ideas.

On the algorithm side, this is also ideal for developing learning-based visual QA algorithms. If we can expand the amount of data on visual responses to diverse contents, distortions types, severities, and temporal behaviors, then we should be able to build much more effective generic, holistic, and distortion-agnostic QA algorithms that operate without reference.

Come to think of it, this might be the answer to my "unlimited resources" question above.

Moderator: These days, each generation has different habits in the digital era. For example, young kids are very familiar with touch screens, which is not the case for older generations. Does this play a role in how quality is perceived?

Al Bovik: "Habits" is the key word here. This implies visual behavior, which as I mentioned above, is key to understanding quality perception. We will naturally find that behavioral models may vary with the interface. I think also in the "sound-bite" generation, temporal duration models may evolve and change significantly. Broadly, the expectations of consumers are for continually increased visual quality and diversity; so models will reflect this trend.

Sebastian Möller: Of course, quality does not only relate to presented media, but also to the interaction involved in obtaining and using the service. Definitely, usability plays a role here, but also the nonfunctional, or hedonic, aspects, like appeal, attractiveness, and joy-of-use. Corresponding metrics are already on the way. Unfortunately, most of the standardization bodies that deal with media quality do not work on interactive systems quality, and vice versa. The overall experience of a multimedia service will depend on both the media quality and the quality-of-use, including hedonic and pragmatic aspects.

Quan Huynh-Thu: I have previously mentioned that context is crucial in quality perception. But another factor is expectation. For a given application, the expectation is de facto related to the reference point that one has in terms of visual quality. With the evolution of the Internet technology and related services, evolution of mobile/computing devices, expectations of users change and therefore reference points also. This may not fundamentally change how visual quality could be modeled (i.e., the structure of the algorithms) but certainly changes the subjective quality benchmark. New devices also bring new ways of interaction, which can impact the way users perceive content. So far, quality assessment has mostly focused on so-called passive scenarios, i.e., where users are asked to just view/listen to content to rate its quality.

Stefan Winkler: Quality is indeed highly dependent on what people are used to and exposed to. I like to compare the times when everybody had VHS players (which were considered perfectly acceptable quality at the time) to today when people have HD TVs and Blu-ray players in their homes, and probably would cringe at the thought of watching a movie on a VHS tape.

Ghassan AlRegib: Let me target both questions above and provide one answer to both. As interactivity is increasingly becoming key in the user's experience and thus in the QoE, defining the quality of interactivity needs more investigation and research to better understand this new world. Researchers in cognitive sciences have recently figured out the "uncanned valley," i.e., why humanoid robots creep us out. The key reason is the mismatch between movement and humanoid traits. In such applications, even if the designed humanoid robot is a top technology, the lack of "realistic" interactivity movements will negatively impact the human perception. Similarly, in interacting with multimedia; if we spend all our efforts on making the multimedia quality top notch while creating an average interactivity experience, the QoE will not score high. In fact, we could use this interactivity quality to give us some space in the multimedia quality and not to demand a high media quality all the time as interactivity might compensate for the low media quality. Of course, more scientific investigation is required here to arrive at these speculations or other conclusions.

I have an App on my iPad that teaches the alphabet to my two-year old daughter. After spending some time on the iPad, she moved to the TV and tried to change the screen by touching it as she did with the iPad. To her, and perhaps, to many other kids, interacting with media is more important than the media itself. This touch-and-feel experience defines quality for them.

Joyce Farrell: Touch screens certainly affect the user's QoE. Gestures, screen size, display temporal response, and finger size are important factors that influence our experience of touch screens. It would be useful to have standardized tests to evaluate the "quality" of a touch screen, where quality is defined by task performance as well as subjective judgments. Expertise also has a very big effect. Perhaps we should study how the experience of a novice changes as he or she gains expertise in the task.

Moderator: What are the application-specific factors in multimedia quality evaluation, i.e., in 3-D displays, medical applications, online and interactive games, and streaming multimedia (e.g., speech, video)?

Joyce Farrell: This is a very good question, one that we should ask about every application. One way to answer this question is to build simulation environments that allow us to manipulate factors that we believe are important for any given application and to determine the impact they have on task performance and/or user experience.

Al Bovik: If I knew the answer to this then I would be writing proposals to quite a few different funding agencies. But seriously, the main factor is modeling. For example, a recent hot topic is the effect of visual quality on recognition (e.g., of faces). Well, to answer this, we need a face model that is useful for the problem and that is perceptually relevant. This relates to my comment on the need for better models of higher-level processing along the neutral stream. In medicine, we'll need models of the organs involved and of

the physics of the imaging modality, for starters.

Ghassan AlRegib: I think such applications will affect the way we "define" multimedia quality and as a result the community will come up with ways to evaluate this newly defined quality. This is where the dimensionality and the complexity of the problem (i.e., the multimedia quality) become much harder than what we are used to. This is where the quality becomes application centric. For example, in social media, the purpose of the shared media is to convey a certain message from the individual sharing the media to her or his peers in the network; the media might be tagged or have a comment associated with it. If the message is in the audio, do we have to define the quality as a function of the audio only and not the video? If the message is in some text embedded in the video, do we pay more attention to the visual data and ignore the audio? Perhaps, if we know the "interest" of the individual, then we might have some idea on what message is being conveyed in a particular media and evaluate the quality accordingly.

Quan Huynh-Thu: I'll focus on 3-D video. In the end-to-end chain of 3-D video delivery, many points can impact quality at signal production/capture, transmission and signal rendering. Concerning the signal production/capture, several factors can impact the quality: geometric distortions/misalignment, color distortions, disparity, and imperfect 2-D-to-3-D conversion. Transmission over error-prone channels will obviously impact quality. Here we also have the headache of keeping the views of the 3-D signal synchronized in case of errors. Concerning the rendering, 3-D displays clearly play an important role in the 3-D QoE. The same stereoscopic 3-D signal displayed on two different 3DTVs may produce different experiences. There is also the adequacy between content capture and display, i.e., ideally 3-D content has to be produced for a given viewing environment (display size and viewing distance). A deviation from the target-viewing environment for which the content has been produced is susceptible to generate distortions such as shape deformation and depth distortion.

Patrick Le Callet: The impact of technology on quality of diagnosis is a seminal issue while considering medical imaging systems such as magnetic resonance imaging and a positron emission tomography scan. The quality of a system relies on its ability to minimize wrong decision of the practicians. This is an extreme case that required full adaptation to the application scenario. Signals are very specific (sometimes the relevant information looks more like noise for nonexperts) and to assess their quality it is needed to consider the pathology under study, the related anatomy, and the image modality. This is part of the medical imaging readers' expertise that a metric should be able to mimic. Of course, it is more realistic to tune a metric for one particular combination of pathology/anatomy/image modality. As all the elements of the chain, from acquisition to visualization, might affect the final decision, it is required to well understand the consequence of technological aspects on the signal while measuring the value of one particular element of this chain.

Moderator: 3-D video quality is an emerging field; what are the challenges ahead?

Al Bovik: What a booger this problem is! First, we lack models of the statistics of the natural 3-D world. We will require this to make good progress. Second, despite 40 years of research of stereopsis and other modes of 3-D, we lack understanding of key perceptual elements of 3-D perception. For example, we do not understand yet how the stereo sense affects the perception of distortions. Does high stereo disparity activity make luminance distortions less visible similar to luminance masking? Recent studies suggest, unexpectedly, that the opposite might be true. Does viewing in 3-D change where we look? Certainly, and unexpectedly, it seems. There are many other examples where our understanding is quite poor. This compounds the fact that 3-D video quality is a double-blind problem: since the perception of the 3-D signal (distorted or otherwise), termed cyclopean signal, occurs only in the brain, we have no way to quantitatively access either the distorted signal or a reference signal. We are thus left with estimating these, before predicting quality. As others have pointed out, 3-D QoE is multifaceted with discomfort, distortion, and display issues (one might call these the three "D"s) all playing a role. However, we poorly understand how each of these relate to overall QoE, and we haven't begun to understand how they combine to affect QoE. We are taking a step-wise approach by examining these issues separately. In particular, I strongly disagree with one of our panelists regarding the success of 3-D quality (distortion) models. We haven't found any that improve upon 2-D models as applied to 3-D data. This remains an open area of research.

Ghassan AlRegib: 3-D video brings interesting challenges to the community on how to evaluate the quality of 3-D videos. Here we are looking at a number of monocular cues from two or more views, and the display is trying to recreate the 3-D world. All steps in the pipeline from acquisition to coding and from streaming to displaying affect the quality of the 3-D video. The impact is usually catastrophic and it results in a great deal of discomfort.

If we consider stereo videos, for example, a slight variation in the color between the two views will result in a discomfort. In depth-based videos, we overcome this problem by using a single reference to generate the second view but this brings it own challenges such as occlusion and disocclusion.

Another challenge is the fact that we do not have a reference in 3-D videos. We capture imperfect views and we try to reproduce 3-D video that has the least artifacts and the most natural views even though the captured views we started with are far from perfect. This introduces challenges, but it also leaves room for us to innovate in processing and displaying the 3-D videos.

From our research at Georgia Tech, we found out that it is a failure to think of the 3-D video as the combination of two 2-D signals. In this case, a few quality measures have been proposed where the

overall quality measure is the combination of individual views quality; for example, the average PSNR of the two views. This oversimplification results in mismatch between measured quality and subject quality. In contrast, one has to consider a 3-D video as a 3-D signal and create a quality measure that is designed for a 3-D signal with certain components.

Finally, the whole R-D analysis for 3-D videos is based on a set of "new" quality metrics that are needed to be designed for 3-D videos. This will open the door for a number of innovative approaches and algorithms.

Quan Huynh-Thu: The concept of 3-D video quality is in fact multidimensional: it includes the signal quality but also other dimensions such as visual comfort and depth quality. These could be termed basic perceptual dimensions. Existing 2-D video quality assessment algorithms are designed to address only the first one (signal quality). 3-D video quality is not simply an extension of 2-D video quality with depth information. Furthermore, even if we consider only the signal quality, there are artifacts that are only specific to 3-D video and not existing in 2-D, so existing 2-D quality assessment algorithms are not designed to handle such 3-D-specific distortions. Second, the subjective assessment of all these dimensions, in particular visual comfort and depth quality, is not easy. Keep in mind that subjective assessment is only meaningful if results are repeatable and reproducible. Currently, it is not clear how to reliably and meaningfully assess subjectively visual comfort and depth sensation/quality of stereoscopic 3-D motion sequences, especially for long durations. Third, in 3-D video, not only the signal itself but also the rendering of the signal can impact significantly the subjective quality. 3-D displays have improved over the years but still suffer from crosstalk, which can decrease the 3-D QoE. Last but not least, in the case of 3-D video, the true signal to assess should in fact be the one reconstructed inside the brain and not the signal displayed on the screen. The bottom line is, if one wants to assess the quality of the 3-D signal as truly perceived, the binocular fusion process should somehow be modeled and integrated—this is not trivial as we need to understand how we actually integrate all the different (monocular and binocular) content cues together.

A paper published at the 2010 IEEE International Conference on Image Processing titled "Video Quality Assessment: From 2-D to 3-D–Challenges and Future Trends" summarizes why 3-D video quality metrics cannot be simple extensions of 2-D video quality metrics and which points in the 3DTV transmission chain are susceptible to affect the QoE at the end-user side.

Patrick Le Callet: As a coauthor with Quan of this paper, I of course fully agree with this vision, with two more comments:

- Lack of ad hoc ways to measure subjectively 3-D QoE is currently a trap for most of quality metric designers. Using standard 2-D quality assessment protocols leads to oversimplification of ongoing 3-D quality metric efforts as pointed out by Ghassan. There are a bunch of 3-D quality metrics in literature that are quite successful to correlate with MOS obtained using standard protocols catching the visual quality. This is just one piece of the puzzle, as other key perceptual dimensions like discomfort and depth sensation are out of this game. I definitively recommend being very cautious while interpreting the results of such metrics. For instance, there is a trend to use asymmetric coding conditions to save bandwidth for one of the view in stereo transmission. In terms of visual quality, it might work as visual perception of most observers is able to align the quality on the "best" view. Nevertheless, this compensation might fire a cognitive load that could have some dramatic effects considering the discomfort dimension. This effect will be transparent for most of usual subjective quality assessment methodologies that uses short video clips, and consequently also for quality metrics tuned on such ground truth.
- 3-D right now means mostly stereo 3-D (S-3-D) (providing two different views to the use whatever the display, shutter, passive, and lens auto stereoscopic displays). It cheats our visual perception enhancing one depth cues: binocular disparities, but we are able to perceive depth from many other cues. Enhancing one of the cues might affect the other ones in a nonreliable way (e.g., losing resolutions, due to interleaved stereo format, affect monocular cues such as texture gradients) that leads to the question of the value of such solution. Considering the challenge ahead, quality assessment should not only focus on S-3-D but should be provide some answers on the values of all candidates technologies such as motion parallax-based ones, holography, etc.

Stefan Winkler: As others have already elaborated, 3-D quality is a highly complex subject with many different issues, which we have only begun to explore and understand.

I'd like to highlight another important aspect: In 3-D it's not just about the best-looking content anymore. Stereoscopic content has potential psychological and physiological effects on a significant portion of the population: if 3-D is not produced, processed, and presented correctly, it can make viewers dizzy or nauseous. Understanding why this happens for some people more than for others, and how these effects can be minimized will be crucial for the success and wide adoption of 3-D.

Joyce Farrell: There are challenges in nearly every aspect of 3-D imaging, including image capture, processing, transmission, and display. But the biggest challenge, it seems to me, is to quantify the value of 3-D when it comes to the users' experience. Industry's fascination with 3-D displays resurfaces every ten years or so. Will this be the year that 3-D displays find their way into people's homes and theatres? For me, this is still an open-ended question. Even if we can solve the technical challenges, will people value the 3-D experience enough to pay for it?

There are at least two applications for which the answer to this question is yes. One application is medical imaging, such as robotic assists in products like the

DaVinci surgical unit. Another application is video games. The impact of 3-D imagery on these applications is undisputable. But for someone who is not a surgeon and who does not play video games, the question of how much I will pay for 3-D imagery remains unanswered.

Moderator: How might the scientific progress and discovered principles in multimedia quality evaluation research benefit other fields?

Quan Huynh-Thu: Research on multimedia quality, and on objective metrics in particular that can predict subjective quality, is a crossroad between several fields such as image processing, computer vision, cognition, neurosciences, and psychology. So the understanding and modeling of how we perceive a signal and form an opinion of quality ultimately benefit the knowledge in all these fields.

Patrick Le Callet: It helps to better tune the technology to the final user. Whatever the multimedia application, adopting a user-centered approach for the development of technologies as a piece of a system of a whole product is much welcomed as it is done for the design of end-user products. The user-centered product design community has certainly developed excellent approaches nevertheless face some difficulties while addressing the lower-level pieces of technologies. For the latter, a good understanding of the underlying technology is mandatory in addition to human factors considerations. This is where our community could be helpful: to fill the gap between engineers, cognitians, and designers.

Stefan Winkler: I don't think we need to look very far for other fields where quality assessment can be beneficial.

As I mentioned earlier, good quality models should be integrated in all kinds of multimedia processing chains for best results, but very few are designed for that. Better quality metrics than PSNR should be used in encoders for rate control, in LCD displays for content preprocessing, in cameras for photo optimization, etc. Unless the quality assessment community can come up with metrics that not only perform well but are also easy to use, easy to interpret, and well accepted, adoption by other communities will be even slower.

Joyce Farrell: Perhaps we should ask how can we make scientific progress and discover principles in multimedia quality evaluation. One approach that I advocate is to build simulation environments that control every aspect of the multimedia signal, including properties of the capture, processing, transmission, and display. In this way, we will be able to determine the critical components that influence subjective quality. Of course, we still have the question about how best to characterize the users' experience.

Clearly, evaluating the user's experience is a challenge. Today, we ask people about their experience after it has already happened. We are initiating a research project at Stanford that will monitor users' brain activity in real time as they view video imagery, both 2-D and 3-D. Whether this produces better information than subjective reports about fatigue or enjoyment is an open-ended but important question to answer.

Al Bovik: Aside from the seeming limitless realm of image/video/multimedia applications that will benefit by using quality models to monitor, assess, and control the quality of visual signal traffic using successful QA models, in my opinion, understanding visual quality and visual distortion perception is fundamental to understanding vision. I view visual distortions and how they are perceived as visual probes into perception, much as visual illusions are. If we come to understand how distortions and their severity are perceived, then we will likely have made significant inroads into understanding a wealth of other visual principles.

Ghassan AlRegib: I think understanding how to evaluate the quality of multimedia will help us understand the source of such complex signals, e.g., human speech, light fields, human vision, etc. This will open the door for many scientific discoveries. For example, if we truly know how the human brain and vision system views things in terms of quality, then we can design better car "visual" systems and better robots.

Moderator: In a recent standardization effort, specification of the hardware the displays was well spelled out. What role does hardware play in affecting how visual quality is perceived? Do we need to come up with a set of quality measures that depend on the hardware?

Sebastian Möller: I think that we should always take into account the full channel, including the sending and receiving (hardware and software) elements. A model that does not explicitly mention the hardware used will make, nonetheless, assumptions about it, specifically from the test situations that have been used for collecting the data material for the model. These assumptions should be made explicit, and care should be taken when interpreting the model predictions in cases where the sending and receiving elements are different.

Patrick Le Callet: Displays are the ultimate steps that translate information into the real world. As long as we are considering mature enough technology, this step can be modeled once for all and being transparent for some metrics. Regarding S-3-D jungle technologies, not only displays but also formats better be cautious and properly define the conditions of the validation of a metric.

For peaky applications such as medical imaging, quality of the display, from a technological point of view, is a strong issue.

There are several standardization efforts to define quality of displays [International Committee for Display Metrology (ICDM), American Association of Physicists in Medicine (AAPM)], and most of them are providing very useful measurements of the technological parameters of displays, sometimes considering human perceptual properties. Image quality assessment, as going toward QoE and targeting more application specific context, should better try to understand the impact of these quality displays factors on the overall visual perceptual quality.

Quan Huynh-Thu: In any standard, the scope is highly important. A technology such as a quality assessment algorithm is designed for a given scope (e.g., types and severity of degradations) and has been validated on subjective data collected in a given scenario, which includes

a certain type of display (specifications). In recent standardization efforts by the ITU supported by extensive work from the Video Quality Experts Group (VQEG) concerning (2-D) video quality metrics (ITU-T J.247, ITU-T J.341), studies have examined whether the display had a significant impact on the subjective quality. It was found that for the given scope of degradations considered in the subjective testing, a very high similarity was found between subjective results even if different displays were used, provided the displays had some minimum performance criteria. So ultimately the standard does not specify any display for which the metric has been validated but de facto a quality assessment metric has been validated on subjective data that has been collected in a certain scenario (including the display that was used to show the videos to the participants). With the current state-of-the-art 2-D display technology, the influence of the display on visual quality is highly dependent on the application. For the types of applications considered in ITU-T J.247 and ITU-T J.341, the display (again with some minimum criteria) did not impact the visual quality (especially for naive viewers) but in other applications, where, for example, color fidelity is crucial or in medical imaging, obviously display is important. Now that said, 3-D video is a different case, as I commented previously. The current 3-D display technology has not reached a level of maturity where the display can be considered to be transparent and clearly there is a need to either define metrics that include the influence of the display or make sure that the display used in subjective testing is "transparent" enough.

Stefan Winkler: Displays clearly play an important role in determining QoE. Perhaps the wider question is, how much different is the experience in an actual home viewing environment from subjective experiments performed in the lab under rigorously controlled conditions? This is not only about the quality and settings of the display, but also about viewing setup, lighting conditions, length of viewing (a whole movie versus a 10-s clip), attention, interest, etc. Of course, we can design quality metrics to work under the "best possible" viewing conditions and displays, but it may be worth exploring how these other parameters can be taken into account.

Joyce Farrell: Quality depends on hardware, software, and the properties of the measurement device, be it a human or an instrument.

At a minimum, we should report the conditions in which we collect subjective judgments of image quality, including the viewing distance, the ambient illumination and the spectral power of the display primaries, the display gamma, and the display resolution. We should also be sure that subjects have corrected to normal vision. And it may be useful to record their age and sex. I hope that in the future, databases that match images and video with mean opinion scores (MOS) will include this information.

And if we want to understand the role that hardware (and software) plays in determining subjective image quality, we need to be able to independently control different hardware (and software) components and record their impact on subjective judgments of image quality. This is why I am a strong proponent of simulation environments.

Moderator: What is the next challenge in the quality arena?

Sebastian Möller: One of the challenges I see is to come to realist predictions of what people do, i.e., how they behave when being confronted with a service. We have invested a lot in making predictions about people's perceptions, but not so much about people's actions. However, when we deal with interactive services, the user's actions will be of paramount importance for the overall quality. What is needed are models that describe user actions both on a semantic (intentional) and on a surface level, i.e., the level of observable actions.

Another challenge I see is to link quality to other aspects affecting the acceptance of a service. One of them is the price or economical benefit. Another is security. People inherently establish tradeoffs between these aspects, and it would be wise to not concentrate on quality alone when designing a new service.

Patrick Le Callet: It appears that with the explosion of applications and technologies, QoE assessment is the key. This is not a new concept: I still remember Touradj Ebrahimi mentioning this emergency in 2001 during a conference keynote, ten years ago already, but we have still far to go to provide satisfying solutions. How do we go beyond? Following

Perreira's triple sensation-perception-emotion user model is a way that tries to measure the following different steps: 1) sensation factors that relate to the sensory perception of (multimodal) content, i.e., human vision, audition, and other sensory perception; 2) perception factors that relate to interpretation of the information from the media, user's satisfaction as cognitive experience (usability, task performance, and information assimilation); and 3) emotion factors that relate to the intensity of emotional experience.

Another way could be to see QoE from a user's experience following the Mahlke model of user experience. As components of user experience, Maklke identifies: 1) the perception of instrumental qualities: usefulness (utility), usability (efficiency, controllability, helpfulness, and learnability); 2) emotional user reactions: subjective feelings, motor expressions, physiological reactions, cognitive appraisals, and behavioral tendencies; and 3) perception of noninstrumental qualities: aesthetic aspects, visual aesthetic, symbolic aspects, associative symbolic, communicative symbolic, motivational aspects, and multimodalities.

Whatever the approaches, QoE assessment will probably occupy researchers for many years. Hopefully, there are some ongoing collaborative actions towards this goal. The European Cooperation in Science and Technology (COST) action network on QoE in multimedia systems and services (QUALINET) is currently gathering the efforts of partners from more than 30 countries toward this goal.

As short-term realistic goals, I see two interesting scopes beside 3-D for quality assessment. The first one is dealing with immersion through two technological aspects: super high resolution and high dynamic range (HDR). The second one is related to multimodality, trying to better understand relationship between visual and haptic perception.

Ghassan AlRegib: In the short term, I think the overall immersive technologies will be heavily investigated. This includes 3-D and interactivity via haptics or via touch. Also, social media will receive quite a bit of attention to determine the quality of a social network based on communicated media and the associated reactions and responses.

Overall, I agree with Patrick, and I believe we are still far from having clear and thorough understanding of how perception, sensation, and emotion interact to define "quality." Add to this the complex pipeline of systems media content undergoes before it is perceived by the user.

Quan Huynh-Thu: Quality measurement alone is not really meaningful per se. In a real-world scenario, its meaning is always linked to a context, which can include the application/service and fees that users must pay. Quality measurement has been so far used mostly for troubleshooting and network/application tuning. Other points that are related to quality are usability and acceptability. Linking these two points to quality measurement is not easy and so application/service dependent that finding a generic way to model their relationship would be quite remarkable. We would almost need the types of models they use in finance.

Stefan Winkler: I see three main challenges: One is new technologies, such as 3DTV, or upcoming display technologies. Another is related to my comments on granularity: We need to find useful applications for quality measurement beyond just monitoring and comparisons, and for that we need to look beyond MOS and overall quality. The third is the rather narrow signal processing focus we usually have in terms of what quality constitutes. Aspects of interaction, ease of use, personal preference, context, emotion, relevance, and appeal, are as important (if not more) than compression and other distortions and need to be taken into account for a comprehensive assessment of the true "QoE."

Joyce Farrell: One challenge is to develop new methods and standards for assessing the user's experience. MOS is clearly limited. Rather than one method, we need multiple methods that quantify different aspects of the quality of multimedia.

Another challenge is to control the complex interactions between multiple factors that influence the user's experience. Video content, image capture, processing, transmission, display, viewing conditions, and user characteristics (expertise, age, sex, vision, etc.) all influence the quality of the multimedia experience. We need ways in which we can independently control different components of a multimedia system as well as methods for analyzing design tradeoffs.

Al Bovik: I believe that as our models improve and become adaptive and intelligent, we should seek to deploy "visual quality agents" in every switch, router, access point, TV, smart phone, camera, and so on. These quality agents should ultimately interact, enabling distributed network control of video traffic, and perceptually optimized acquisition, transmission, coding, compression, and display of visual information. This implies huge deployments, and yes, I think the problems we are working on are this important.

MODERATOR

Fatih Porikli (fatih@merl.com) is a Distinguished Scientist at Mitsubishi Electric Research Labs (MERL), Cambridge, Massachusetts, United States. He received his Ph.D. degree from the Polytechnic Institute of New York University. Before joining MERL in 2000, he developed satellite imaging solutions at Hughes Research Labs and 3-D capture and display systems at AT&T Research Labs. His work covers areas including computer vision, machine learning, compressive sensing, sparse reconstruction, video surveillance, multimedia denoising, biomedical vision, radar signal processing, and online learning. He received the 2006 R&D100 Award in the "Scientist of the Year" category (a select group of winners) in addition to numerous best paper and professional awards. He serves as an associate editor for many IEEE, Springer, and SIAM journals. He was the general chair of the 2010 IEEE International Conference on Advanced Video and Signal-Based Surveillance and organizer of several other IEEE conferences.

PANELISTS

Al Bovik (bovik@ece.utexas.edu) holds the Curry/Cullen Trust Endowed Chair professorship at The University of Texas

at Austin, where he directs the Laboratory for Image and Video Engineering (LIVE) in the Department of Electrical and Computer Engineering and the Institute for Neurosciences. He is broadly interested in image processing and modeling of visual perception. He has received many awards for his work including most recently the 2011 IS&T Imaging Scientist of the Year Award and the 2009 IEEE Signal Processing Society Best Paper Award. He is also known for helping to create the IEEE International Conference on Image Processing (ICIP) and cofounding *IEEE Transactions on Image Processing*.

Chris Plack (Chris.Plack@manchester.ac.uk) is the Ellis Llwyd Jones Professor of Audiology at the University of Manchester. He specializes in human auditory perception, in particular, the physiological mechanisms that underlie our perception of loudness and pitch, and the effects of hearing loss on these mechanisms. He is a fellow of the Acoustical Society of America, and an associate editor of *Journal of the Acoustical Society of America*. He is the author of *The Sense of Hearing* (Psychology Press) and has edited two volumes: *Pitch: Neural Coding and Perception* (Springer) and *Hearing* (Oxford).

Ghassan AlRegib (alregib@gatech.edu) is an associate professor at the School of Electrical and Computer Engineering at the Georgia Institute of Technology in Atlanta. He is the director of the Multimedia and Sensors Lab at Georgia Tech. He received the 2008 Outstanding Junior Faculty Award. He is conducting research in the area of multimedia processing with recent focus on 3-D videos quality, processing, and compression. He is the area editor for columns and forums in *IEEE Signal Processing Magazine* and is the editor-in-chief of *ICST Transactions on Immersive Telecommunications*. He is the cochair of the IEEE Multimedia Communications Technical Committee Interest Group on 3-D rendering, processing, and communications. He is an IEEE Senior Member.

Joyce Farrell (joyce_farrell@stanford.edu) is a senior research associate in the Department of Electrical Engineering at Stanford University. She is also the executive director of the Stanford Center for Image Systems Engineering. Prior to joining Stanford University, she worked at a variety of companies and institutions, including the NASA Ames Research Center, New York University, the Xerox Palo Alto Research Center, Hewlett Packard Laboratories, and Shutterfly. She is also the chief executive officer and founder of ImagEval Consulting, LLC.

Patrick Le Callet (patrick.lecallet@univ-nantes.fr) is professor at Polytech Nantes-Université de Nantes and head of the Image and Video Communication group at CNRS IRCCyN. He is mostly engaged in research dealing with human vision modeling and its application in image and video processing with the current center of interest in color and 3-D image perception, visual attention modeling, video, and 3-D quality assessment. He serves in the VQEG where he cochairs the Joint Effort Group and 3DTV activities. He is the French national representative of the European COST action IC1003 QUALINET on QoE of multimedia service in which he leads the working group on mechanisms of human perception.

Quan Huynh-Thu (qht@ieee.org) is currently a senior scientist at Technicolor Research & Innovation. His main research interests for the past ten years have been video quality assessment, human factors, and visual attention. He holds the Dipl.-Ing. degree in electrical engineering from the University of Liège (Belgium), the M.Eng. degree in electronics engineering from the University of Electroc-Communications (Japan), and the Ph.D. degree in electronic systems engineering from the University of Essex (United Kingdom). He codeveloped a perceptual video quality metric included in the ITU-T Recommendation J.247 for the objective measurement of video quality. He is a rapporteur for Question 2 in ITU-T SG9 and cochair of both the VQEG 3DTV and multimedia groups.

Sebastian Möller (sebastian.moeller@telekom.de) studied electrical engineering in Bochum (Germany), Orléans (France), and Bologna (Italy). He received his Ph.D. degree from Ruhr-Universität Bochum in 1999, and his qualification to become a professor (venia legendi) in 2003. Since 2007, he has been a professor for quality and usability at Deutsche Telekom Labs, TU Berlin, and works on speech and audio-visual quality, spoken dialogue systems, usability, user modeling, and usable security. He is currently acting as a rapporteur for ITU-T Q.8/12.

Stefan Winkler (stefan.winkler@adsc.com.sg) is a principal scientist at the Advanced Digital Sciences Center of the University of Illinois at Urbana Champaign in Singapore. He is also a scientific advisor to Cheetah Technologies. He holds a Ph.D. degree from the Ecole Polytechnique Fédérale de Lausanne, Switzerland, and an M.Eng./B.Eng. degree from the Technische Universitat Wien, Austria. He has published more than 70 papers and is the author of *Digital Video Quality* (Wiley). [SP]

dates AHEAD

Please send calendar submissions to:
Dates Ahead, c/o Jessica Barragué,
IEEE Signal Processing Magazine
445 Hoes Lane,
Piscataway, NJ 08855 USA,
e-mail: j.barrague@ieee.org
(Colored conference title indicates SP-sponsored conference.)

2011

NOVEMBER

2011 Conference on Design and Architectures for Signal and Image Processing (DASIP)
2–4 November, Tampere, Finland.
General Chairs: Tapani Ahonen and Jari Nurmi
URL: www.ecsi.org/dasip

2011 IEEE International Workshop on Information Forensics and Security (WIFS)
16–19 November, Foz do Iguaca, Brazil.
General Chairs: Dinei Florêncio, Nasir Memon, and Anderson Rocha
URL: http://wifs11.org/Index.aspx

2011 7th International Conference on Natural Language Processing and Knowledge Engineering
27–29 November, Tokushima. Japan.
URL: http://aia-i.com/nlpke11/home.php

DECEMBER

2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)
4–6 December, San Antonio, Texas.
General Chair: Yidong Chen
URL: http://compgenomics.cbi.utsa.edu/gensips11/

2011 Automatic Speech Recognition and Understanding Workshop (ASRU 2011)
11–15 December, Hawaii.
General Chairs: Michael Picheny and David Nahamoo
URL: http://www.asru2011.org/

Digital Object Identifier 10.1109/MSP.2011.942529
Date of publication: 1 November 2011

Fourth International Workshop on Computation Advances in Multisensor Adaptive Processing (CAMSAP)
13–16 December, San Juan, Puerto Rico.
General Cochairs: Aleksandar Dogandzic and Maria Sabrina Greco
URL: http://www.conference.iet.unipi.it/camsap11/

IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)
14–17 December, Bilbao, Spain.
General Cochairs: Dimitrios Serpanos and Adel Elmaghraby
URL: http://www.isspit.org/isspit/2011/

2012

JANUARY

IEEE International Conference on Emerging Signal Processing (ESPA)
12–14 January, Las Vegas, Nevada.
General Chair: Panos Papamichalis
URL: http://www.ieee-espa.org/

FEBRUARY

2012 International Conference on Pattern Recognition Applications and Methods (ICPRAM)
6–8 February, Algarve, Portugal.
URL: http://www.icpram.org

MARCH

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
25–30 March, Kyoto, Japan.
General Chairs: Hideaki Sakai and Takao Nishitani
URL: www.icassp2012.com/

MAY

IEEE International Symposium on Biomedical Imaging (ISBI)
2–5 May, Barcelona, Spain.
General Chairs: Alejandro Frangi and Andrés Santos
URL: http://www.biomedicalimaging.org/p

JUNE

IEEE 13th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)
17–20 June, Çesme, Turkey.
General Cochairs: Hakan Deliç and Georgios B. Giannakis
URL: http://www.spawc2012.org

Seventh IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM 2012)
17–20 June, Hoboken, New Jersey.
General Cochairs: Hongbin Li and Xiaodong Wang
URL: http://www.stevens.edu/sam2012/

JULY

2012 IEEE International Conference on Multimedia and Expo (ICME)
9–13 July, Melbourne, Australia.
General Chairs: Jian Zhang, Dan Schonfeld, and David Dagan Feng
URL: http://www.icme2012.org/index.php

SEPTEMBER

2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)
23–26 September, Santandar, Spain.

IEEE International Conference on Image Processing (ICIP)
27 September–6 October, Orlando, Florida.
General Chair: Eli Saber
URL: http://icip2012.com

2013

MAY

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)
26–30 May, Vancouver, Canada.
General Chairs: Li Deng and Rabab Ward
URL: http://www.icassp2013.com/

2011 Index
IEEE Signal Processing Magazine
Vol. 28

This index covers all technical items — papers, correspondence, reviews, etc.— that appeared in this periodical during 2011, and items from previous years that were commented upon or corrected in 2011. Departments and other items may also be covered if they have been judged to have archival value.

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of the paper or other item, and its location, specified by the publication abbreviation, year, month, and inclusive pagination. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication abbreviation, month, and year, and inclusive pages. Note that the item title is found only under the primary entry in the Author Index.

AUTHOR INDEX

A

B

C

D

Digital Object Identifier 10.1109/MSP.2011.942863
Date of publication: 1 November 2011

SUBJECT INDEX

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

D

E

+ Check author entry for coauthors

+ Check author entry for coauthors

N

O

P

+ Check author entry for coauthors

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

Q

R

S

+ Check author entry for coauthors

T

+ Check author entry for coauthors

IEEE SignalProcessing Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page Qmags THE WORLD'S NEWSSTAND®

 (continued from page 200)

physically accurate model of image formation with a reasonably small number of parameters. Regularized inverse methods must not only preserve, but also precisely reconstruct biological structures in their shapes and intensity level. This appears to be an area where nonlinear techniques (e.g., $\ell 1$-norm minimization) hold great promises.

- *Quantitative image analysis (3-D + time)*: Biologists are in great need of quantitative data that can be collected, stored, and subjected to statistical analyses. Concretely, this requires the detection of fluorescent probes, segmentation, particle tracking, shape and motility analysis of cells, and the extraction of gene expression profiles. The challenge there is to develop global algorithms that can address the detection and tracking problems simultaneously, and also handle huge amounts of data, including very crowed image scenes, both in two-dimensional (2-D) and 3-D.
- *Novel methods for the extraction and characterization of 3-D features and structure of networks*: Biological structures are inherently 3-D. They can also be rather complex; for example, the dense networks of neurons, microvascularization, or cellular scaffolds. This calls for the development of a novel panoply of 3-D detectors for finding filaments, sheets, open curves defining singularities without jump in the intensity function, and more generally, key points that are specific to biology. One possible venue is the design of optimized templates using 3-D steerable filters and wavelets or the design of new variational functionals involving appropriate differential operators. It is necessary to develop graph-based models or other mathematical tools to link such structures (missing data managing) and to compose the underlying networks. One also needs to define new dedicated metrics that allow the comparison, classification and assessment of such higher-level entities.

These challenges involve a multidisciplinary context: biologists (preparation, imaging protocol, and interpretation of results), biochemists (staining, fluorescent markers), imaging scientists (design of cutting-edge instrumentation), and finally, image processing. A last aspect that should not be overlooked by signal processors is the design of user-friendly imaging software (e.g., in the form of a plug-in for ImageJ or Icy [4]) that can be readily used by biologists with minimal knowledge of computer science. Ultimately, the best impact that we can hope for is that biologists and microscopists rely on our algorithms to perform their work.

MEDICAL IMAGING

Advances in image acquisition have sparked a steady expansion of the dimensionality of data sets, especially when data sets include local directional information on fiber orientation or four-dimensional (4-D) blood flow. Such high-dimensional data sets cannot be interpreted visually without adequate image analysis and, as such, there is a major ongoing effort toward developing analysis techniques for such complexly structured data.

A second recent trend is the advent of in-vivo molecular imaging modalities such as fluorescence and bioluminescence imaging, which enable the live imaging of gene expression and protein interactions. Combined with detailed structural imaging modalities such as magnetic resonance (MR), the biochemical onset of disease and therapy can be monitored in combination with structural and functional consequences over time. Main analysis challenges are the fusion of heterogeneous data (e.g., registration of photographs to 3-D data of animals with highly variable posture), and linking in-vivo imaging to existing genomics databases, such as the Allen brain database [5].

A third important trend is longitudinal image analysis. Increasingly, imaging data is acquired in a follow-up manner, where the subtle changes over time reflect the information of interest. Longitudinal imaging studies are typically performed to characterize diseases in single patients or patient cohorts. However, large follow-up population imaging studies are also being performed on cohorts of 6,000–20,000 healthy subjects to image the process of healthy aging, and to identify early disease biomarkers upon the onset of disease in individual study subjects [6]. Analysis challenges lie in the sheer data volume that necessitates full automation, and in feature extraction and data mining in such large image databases

Finally, there is a strong trend toward objective quantitative benchmarking of analysis methods, where data sets with expert annotations are made publicly available, and standardized evaluation pipelines can be applied for algorithm benchmarking. Since the initial MICCAI Segmentation Challenge in 2007, competitive challenge events have been organized on, among others, liver, coronary, brain, and carotid segmentation as well as on image registration. Their importance is underscored by the fact that novel analysis techniques for these problems should today often be benchmarked against these public evaluation frameworks to qualify for publication acceptance.

AUTHORS

Jean-Christophe Olivo-Marin (jcolivo@pasteur.fr) is a research director at Institut Pasteur in Paris, France.

Michael Unser (michael.unser@epfl.ch) is a professor at Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

Laure Blanc-Féraud (Laure.Blanc_Feraud@inria.fr) is a research director at CNRS-I3S in Sophia Antipolis, France.

Andrew Laine (laine@columbia.edu) is a professor at Columbia University, New York.

Boudewijn Lelieveldt (B.P.F.Lelieveldt@lumc.nl) is a professor at Leiden University and at Delft University of Technology, The Netherlands.

REFERENCES

[1] "Special section on bioimaging," *IEEE Signal Processing Mag.*, vol. 23, no. 3, May 2006.

[2] "Issue on genomic and proteomic signal processing," *IEEE J. Select. Topics Signal Processing*, vol. 2, no. 3, June 2003.

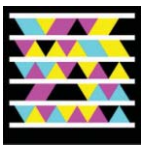
[3] (2003, Sept.). Nature cell biology, suppl. [Online]. Available: www.nature.com/focus/cellbioimaging/index.html

[4] [Online]. Available: http://rsbweb.nih.gov/ij/ and http://icy.bioimageanalysis.org/

[5] [Online]. Available: http://www.alleninstitute.org

[6] [Online]. Available: http://www.mesa-nhlbi.org/ and http://www.populationimaging.eu

Video



Mohammad M. Mansour,
Liang-Gee Chen, and Wonyong Sung

Trends in Design and Implementation of Signal Processing Systems

The design and implementation of SP systems (DISPS) includes the development of software tools and methodologies to support the design of these complex systems. In its early days, DISPS focused on the hardware-based design of SP algorithms to meet real-time requirements. This topic was often called very large-scale integration (VLSI) SP. The emphasis has gradually shifted to include software and hardware/software codesign and implementation aspects. Programmable digital signal processors (DSPs) and embedded central processing units (CPUs) are now popular for real-time SP, such as mobile phones. Field programmable gate array (FPGA)-based designs are also replacing application-specific integrated circuits (ASICs) in many applications. DISPS is now rapidly expanding to include multicore and many-core platforms for throughput demanding applications, low-power SP circuits for power-sensitive applications, and distributed implementations of emerging bioinspired and cognitive radio networks. This column gives an overview of these trends in DISPS as presented during the expert summary, which was organized by the DISPS Technical Committee at ICASSP 2011.

HIGH LEVEL ARCHITECTURE EXPLORATION AND SYSTEM OPTIMIZATION

As the complexity of SP systems continues to grow, it has become too difficult to design and optimize complex system on chips (SoCs) for power-performance-cost while meeting time-to-market constraints. Design methodologies developed for the single-core SoC era are no longer adequate for the task of designing explicitly parallel SoCs. A wide range of architectures exist from traditional DSPs/ASICs/FPGAs to multicore/many-core processors. Optimizing such complex systems is a daunting task. New system design methodologies are needed for multilevel hardware/software co-optimization that enable flexible architecture explorations, extract more parallelism, and optimize global power consumption.

MASSIVELY PARALLEL SP ARCHITECTURES

Advances in integrated circuits (ICs) technology have opened up the possibility of implementing explicitly parallel architectures on homogeneous multicore processors or many-core hardware accelerators, such as graphics processing units (GPUs). Homogeneously parallel applications such as basic filtering, matrix operations, IPv4 packet routing, and image processing are well serviced by these homogeneous architectures. However, increased power consumption, insufficient memory bandwidth, and nonparallelizable algorithms have become the performance-limiting factors of these architectures, and hence new efforts in algorithm and library developments are much needed. Example applications and their performance-critical subsystems include: medical imaging requiring image registration with feature detection and matching, particle filtering for real-time video object tracking, and network processing for higher-level routing decisions requiring partial packet stream reassembly. Challenges also include designing and optimizing massively parallel architectures across different granularities and different dataflow structures.

PARALLEL HETEROGENEOUS SP ARCHITECTURES ON 3-D ICs

Modern and emerging SP applications typically have some components that are not homogeneously parallel. Such applications have parallelism that exists at different levels of granularity with various forms of dataflow structure. They also need to be optimized for latency, throughput and power consumption. Parallel heterogeneous architectures are well suited for these applications; however they demand assembly techniques beyond the boundary of 2-D IC fabrication. Research now is focused on 3-D IC technology to improve performance and break the power-bandwidth barriers by exploiting the extra vertical dimension. Vertical stacking increases transistor density per area footprint, enables integration of many heterogeneous cores and memory blocks, provides massive memory bandwidth, and reduces on-chip traffic delay using shorter interconnect wires in the form of 3-D networks-on-chip (NoC).

FPGA-BASED DSP SYSTEM DESIGN

FPGAs are convenient for fast prototyping, but are considered inefficient in terms of gate density and power consumption. However, as the nonrecurring engineering cost of VLSI fabrication keeps increasing, FPGAs are now replacing many ASIC-based designs even for nonprototyping applications. Specifically, the development of platform FPGAs with hardware multipliers and memory blocks has prompted the use of FPGAs in DSP applications. FPGA vendors now provide not only high-density devices with thousands of on-chip multipliers but also DSP system design environments, such as the

Digital Object Identifier 10.1109/MSP.2011.942318
Date of publication: 1 November 2011

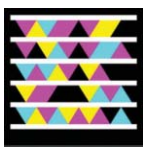
1053-5888/11/$26.00©2011IEEE

system generator of Xilinx. Many FPGA chips also contain embedded processors rendering them complete platforms for DSP system design.

SP AND ERROR CORRECTION FOR NONVOLATILE MEMORY DEVICES

Nonvolatile storage devices in the form of NAND flash memories and solid-state drives have become the storage techniques of choice in many mobile and portable devices. The continued density growth of these devices has been mainly driven by aggressive technology scaling and the use of multilevel per-cell techniques. However, bit errors are becoming more severe as memory process technology scales down below 40 nm. Error-control coding techniques have been employed to improve the endurance and performance of NAND flash memories. However, traditional error correction codes [BCH/Reed-Solomon (BCH/RS)] suffer from increased overhead in coding redundancy and read latency as the number of errors increases. In addition, the number of electrons stored in a memory cell is decreasing with every generation of flash memory resulting in weak signals that require enhanced sensing techniques. Research challenges include reduced complexity coding, enhanced threshold sensing, and adaptive interference canceling techniques.

Slides



DSP-ASSISTED ANALOG AND RF CIRCUITS

As complementary metal–oxide–semiconductor (CMOS) process technology keeps shrinking, the analog portion of SoCs is increasingly dominating the silicon area and power consumption because analog components do not scale well with Moore's law, as does digital logic. Traditional matching techniques that compensate for process variations do not work well as CMOS feature size scales down. Thus, to enhance performance and reduce power consumption of analog/RF circuits, it is necessary to utilize DSP techniques that can be realized with essentially free digital logic. Examples include very high performance analog-to-digital convertors with self-calibration, RF power amplifier linearization, and intermediate-frequency sampling receivers.

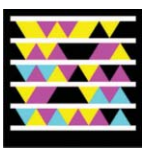
AUTHORS

Mohammad M. Mansour (mmansour@aub.edu.lb) is an associate professor at the American University of Beirut.

Liang-Gee Chen (lgchen@cc.ee.ntu.edu.tw) is a professor at National Taiwan University.

Wonyong Sung (wysung@snu.ac.kr) is a professor at Seoul National University.

Video



Tülay Adalı, David J. Miller,
Konstantinos I. Diamantaras, and Jan Larsen

Trends in Machine Learning for Signal Processing

By putting the accent on learning from the data and the environment, the Machine Learning for SP (MLSP) Technical Committee (TC) provides the essential bridge between the machine learning and SP communities. While the emphasis in MLSP is on learning and data-driven approaches, SP defines the main applications of interest, and thus the constraints and requirements on solutions, which include computational efficiency, online adaptation, and learning with limited supervision/reference data. While MLSP has always been an active area, it is now converging toward the very center of activity in SP research due primarily to two underlying reasons:

- As data now come in a multitude of forms and natures, it has become evident that solutions must emphasize both learning from the data and minimizing unjustified assumptions about the data generation mechanism. Simplifying assumptions such as Gaussianity, stationarity, and circularity can no longer be easily justified, and nonlinearity plays a more important role in today's problems.
- Almost all of the new application areas in SP emphasize the importance of interdisciplinary research, i.e., the need both to work closely with the target application domain and discipline and the need to leverage suitable ideas and tools from diverse disciplines, to develop the best solutions. Many new applications are also sufficiently complex that they require use of multiple, interacting tools and they may need to meet multiple, simultaneous objectives—e.g., both signal classification and signal enhancement, or simultaneous classification and biomarker discovery in medical applications.

Indeed, these two aspects are at the heart of MLSP research. Since learning is emphasized, MLSP methods have always been more data driven than

Digital Object Identifier 10.1109/MSP.2011.942319
Date of publication: 1 November 2011

IEEE SignalProcessing | Previous Page | Contents | Zoom in | Zoom out | Front Cover | Search Issue | Next Page | Qmags THE WORLD'S NEWSSTAND®

model driven; however, the "integrative" nature of MLSP research has also always been emphasized—whenever available, reliable domain information has been integrated, in a principled (e.g., a Bayesian) fashion, into the solutions. Thus, now besides being attractive for many of the traditional SP applications such as pattern recognition, speech, audio, and video processing, MLSP techniques are the primary candidates for a new wave of emerging applications such as brain-computer interface, multimodal data fusion and processing, behavior and emotion recognition, and learning in environments such as social networks and dynamic networks in general.

In what follows, we first discuss current areas of significant activity as well as emerging trends, first in terms of theory, and then applications. Specifically, we discuss the trends in learning theory, and in particular, discuss a major paradigm shift in learning as demonstrated by cognitive information processing. Then, we discuss the role MLSP plays in a number of key emerging application areas.

TRENDS IN LEARNING THEORY

In terms of theory, graphical and kernel methods, Bayesian learning, information-theoretic learning, and sequential learning have always been important areas of activity within MLSP. The need for nonlinear adaptive algorithms for advanced SP and streaming databases has fueled interest in a number of areas, including sequential active learning, which includes as an important subclass: kernel adaptive filters. Sequential learning algorithms are a fundamental tool in adaptive SP and intelligent learning systems as they embody an efficient compromise among constraints such as accuracy, algorithmic simplicity, robustness, low latency, and fast implementation. In addition, by defining an instantaneous information measure on observations, kernel adaptive filters are able to "actively" select training data in online learning scenarios. This active learning mechanism provides a principled framework for knowledge discovery, redundancy removal, and anomaly detection.

DISTRIBUTED LEARNING

With ever-growing data set sizes in real-world applications involving petabytes of information, it is becoming increasingly important to distribute learning tasks by assigning subsets of data to different processors. The processors thus need to communicate and exchange information in such a way that the overall system collectively solves the problem in an optimal manner. There is a wide range of such applications, including multiple agent coordination, estimation and classification/detection problems in sensor networks, and packet routing problems, among many others. The development of learning algorithms for distributed and/or cooperative scenarios, where several nodes have to solve the same or similar classification/estimation/clustering tasks, is therefore becoming an important area of increasing interest within the machine learning community. Typically, algorithms that work in these environments need to conform to limitations in data sharing among the nodes due to either energy/bandwidth constraints or privacy issues. Related applications that are inherently distributed include sensor networks problems and learning in social networks.

SPARSITY-AWARE LEARNING

Sparsity is a natural property of many systems of interest in SP, and sparsity-aware systems have been shown to offer improved performance over sparsity-agnostic ones. Accordingly, this is a topic of growing interest. In mobile communications, for example, MLSP methods can exploit the sparsity present in the network to improve the estimation of channel parameters and timing delays. Sparsity can be due to user inactivity, to the structure of the channel, or to the network topology.

SEMISUPERVISED LEARNING

Numerous SP applications, both traditional and de novo, involve classification and detection, e.g., various speech recognition tasks, music genre classification, entity recognition in video, emotion detection, and network traffic classification based on packet time series, to name just a few. Traditionally, these statistical classification applications have been treated as supervised learning problems, with the classifier designed using a training set of supervised (labeled) examples. However, in many domains, given pervasive sensing and massive data storage capabilities as well as large publicly accessible data repositories, it is both easy and inexpensive to collect a huge "training set" of examples; on the other hand, ground-truth labeling them is both enormously time-consuming as well as expensive, depending on the domain. This labeled/unlabeled data asymmetry motivates semisupervised learning techniques, which generally aim to enhance the (poor) classification performance achievable using a small (deficient) labeled training set by leveraging, for training purposes, many unlabeled samples.

Semisupervised techniques are either generative—modeling the joint density of the feature vector and class label—or discriminative—focusing solely on optimizing the class decision boundary. They may perform either inductive inference—imposing an explicit decision boundary on the feature space—or transductive inference, where the test set itself is effectively treated as part of the unlabeled set, used for joint semisupervised learning and inference. They may assume identical training and test set class distributions. On the other hand, an important recent trend is domain adaptation, where the test set distributions may be different and, thus, where recalibration of the classifier for the test set, albeit an ill-posed problem, may be required. Here a new (test) "domain'' may imply a new sensing environment, i.e., changes in where or when data sensing occurs, relative to training. This may also correspond, e.g., to applying a speech recognition system trained on one population to a different population or a network traffic classifier trained on one local network to a different one. Semisupervised domain adaptation is a ubiquitous problem, with many potential (application-specific) factors that may contribute to statistical differences between the training and test domains.

An underlying theme in many recent MLSP approaches is the need to

deal with multiple objectives, i.e., to deemphasize the traditional optimality with respect to a single chosen metric. In addition to the focus on the traditional bias-variance dilemma—always emphasized in MLSP research so that methods will generalize well to unseen data—the set of objectives now also includes robustness, efficiency, and full interaction with the environment for a complete (global) performance assessment. All these considerations, among others, define the cognitive information processing paradigm, which is discussed next.

COGNITIVE INFORMATION PROCESSING

Artificial cognitive systems and cognitive information processing are emerging trends and will play an increasingly important role in MLSP in the coming years. The ability to perform cognitive information processing can be seen as a natural progression of MLSP, aiming to revitalize some of the original ideas of Alan Turing's "Theory of Computation" and Nobert Wiener's "Cybernetics" and those subsequently pioneered in the SP community by Bernard Widrow. The grand vision is to design and implement profound cognitive information processing systems for augmented human cognition in real-life environments. The practical imperative of this vision is driven by global megatrends related to pervasive and distributed computation, connectedness of people and systems, and pervasive digital sensing, which just a decade ago would have been impossible.

Cognitive information processing (CIP) involves the ability to perceive, learn, reason, and interact robustly in open-ended changing environments by integrating all available information—from multiple raw information sources and sensor inputs to user-driven feedback, annotations, and descriptions. We suggest using a tiered description of cognitive functionality: from low-level, simple sensing-action processing to high-level processing such as decision making and goal planning. There have been other suggestions setting out a minimal set of conditions for signifying processing as being "cognitive." In Simon Haykin's formulation, a cognitive information processing system would require the presence of four properties: 1) Perception-action cycle processing; 2) memory, to predict consequences in the environment; 3) an attention mechanism for allocation of resources; and 4) intelligence/reasoning for decision making in uncertain and complex environments. The important discussion aiming to fully formalize a definition of cognitive information processing and cognitive systems is ongoing; however, many concrete models, systems, and engineering solutions are already emerging.

Machine learning models that continuously learn from both data and previous knowledge will play an increasingly important and instrumental role in all levels of cognition in the real digital world that consists of large data sets, complex, distributed, interacting systems, and unknown, nonstationary environments—this is all usually too complex to be modeled within a limited set of predefined specifications. In real-life environments , there will be inevitably a need for CIP-based automated robust decisions and behaviors in novel situations, including the handling of conflicts and ambiguities. Hence there is a quest for dynamical learning systems, that continuously adapt to changing environments—one of the central components of machine learning for SP. Further, there is a need, beyond capabilities of current systems with built-in semantic representations, for automatic extraction and organization of meaning, purpose, and intention in interplay with the environment and with entities that include computers, embodied agents (i.e., humanlike artificial systems), and human users. In particular, interactive user systems (users-in-the-loop) models will be of vital importance.

Current examples of the use of machine learning in cognitive information processing include e.g., cognitive radio, personalized information systems, sensor network systems, social dynamics systems, semantic analysis systems, Web 2.0 and beyond, and cognitive components analysis. It is also obvious that the success of such approaches requires a multidisciplinary team effort with mixed competencies in engineering, computer science, statistics, machine learning, psychology, neuroscience, and specific domain knowledge.

TRENDS IN MLSP APPLICATIONS

The integrative nature of MLSP techniques has made them primary candidates for many of the emerging applications—a long list that includes brain-computer interface, behavior and emotion recognition, and learning in environments such as social networks and dynamic networks. Next, we discuss three such applications that have received particular attention within the community.

MULTISET DATA ANALYSIS AND MULTIMODALITY DATA FUSION

Analysis of multiple sets of data, either of the same type—multisubject data, data measured at different (time, space) points or under different environments—or of different types, as in multimodality data (e.g., audio and video, or different medical imaging data) is inherent to many problems in SP.

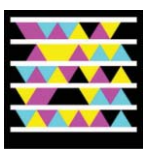
A good example is biomedical image analysis, which is especially challenging because of the rich nature of the data made available by various imaging modalities. Many biomedical studies collect multiple data sets, such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG), structural MRI (sMRI), and genetic data, in addition to clinical and behavioral data and other subject-based assessment parameters. Efficient use of all this information for inference, while minimizing assumptions made about the underlying nature of the data and relationships, is a difficult task, but one that promises significant gains, both scientific and, in the long run, societal, for challenging and important problems, such as the understanding of the human brain function. The need to

minimize strong modeling assumptions is especially evident when studying brain function in natural states such as rest, or when performing tasks such as driving. Data-driven methods such as blind source separation and independent component analysis (ICA), which make minimal modeling assumptions on the data and the underlying processes, are particularly attractive in this context as they can achieve useful decompositions of the multimodal or multiset data without strong assumptions, and can also incorporate reliable prior information whenever available. Along with ICA-based techniques, other latent variable analysis techniques such as tensor decomposition are providing valuable tools for the analysis of multiset data and for fusion of multimodality information.

AUDIO AND MUSIC PROCESSING

Audio SP has always been a central part of SP, with many applications, ranging from sound recording and reproduction systems to advanced speech recognition. Machine learning has also been a central component when it comes to understanding and extracting audio information, even in spite of the fact that machine learning models and algorithms have often been developed without any special attention given to physical modeling of the production mechanisms for audio signals. Online streaming and distributions of audio, and in particular music, has opened a new avenue of possibilities for systems that enable interpretation (semantic audio), music organization, interaction, and sharing. This has indeed already revolutionized the way we consume music and in fact has created new global market opportunities. The special issue on music SP in *IEEE Journal of Selected Topics in Signal Processing* (fall 2011) set the stage for current activities in this field. A key component will be the interplay of SP, which enables the extraction of relevant features, and machine learning, which assists with interpretation and representation of results to users.

Slides



SYSTEMS BIOLOGY

At the beginning of the 21st century, the establishment of high-throughput screening methods and the completion of the human genome mapping marked the beginning of a new era for biological research. The contribution of SP to the acquisition and analysis of these data was crucial since biomolecular signals—e.g., microarray-based gene expression profiles, protein spots in gels, mass spectra, biomolecular images—had to be filtered, accurately detected, normalized, and analyzed. In addition to the SP, machine learning methods also began to be used to unravel the biological meaning of the signals and to categorize the evidence in meaningful ways. It was at that stage that well-established supervised and unsupervised machine learning methods started to provide useful answers to even clinically relevant questions, such as finding gene expression profiles that could be used as biomarkers for certain types of leukemia.

Today, about a decade later, despite the continual development of high throughput methods, producing terabytes of data on a daily basis, many important biological questions still cannot be well addressed, and it is widely accepted that bioinformatics data analysis alone is not sufficient to capture the dynamics and emerging properties of living cells, tissues, and organisms. A new field is thus rapidly emerging, that of systems biology, with the main objective to integrate all qualitative and quantitative biological knowledge, extracted either by biological research or analysis, within holistic and useful models that can capture biological system dynamics at different scales (cell, tissue, whole organism) but also across multiple scales.

Hence, although grounded in biology, biochemistry, and mathematics, the contribution of informatics and especially machine learning in systems biology is more requisite than ever. To build integrative dynamical models while extracting the network of pairwise interactions among molecular species and their possibly causal relations, we need powerful MLSP methods to discover as many true interactions as available data sets (of given size) may allow. A challenging problem systems biology is facing today in many different contexts is the joint estimation of parameters and model structures from sparse and noisy time-course data. Although some of the most sophisticated inference methods have already been tried, we are not yet able to train models of sufficient size, commensurate with the (large) number of molecular units. Many derived models thus have to be tuned either exclusively manually or only partially through parameter estimation methods. It is evident that more efficient MLSP methods are needed to learn, from the available data, dynamical system models of high complexity and accuracy, able to be used in realistic simulations for in-silico experimentation, leading to formation of new hypotheses that could drive new biological research.

ACKNOWLEDGMENTS

We would like to thank all the past and present members of the MLSP/NNSP TC who provided feedback and participated in the discussion on "Trends in MLSP." In particular, we would like to acknowledge the feedback from Sergios Theodoridis, Elias Manolakos (on systems biology), Jeronimo Arenas-Garcia, Gustavo Camps-Valls, and Ignacio Santamaria.

AUTHORS

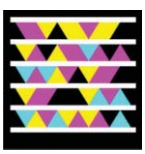
Tülay Adalı (adali@umbc.edu) is a professor at the University of Maryland, Baltimore County.

David J. Miller (djmiller@engr.psu.edu) is a professor at the Pennsylvania State University.

Konstantinos I. Dimantaras (kdiamant@it.teithe.gr) is a professor at the TEI of Thessaloniki.

Jan Larsen (jl@imm.dtu.dk) is an associate professor at the Technical University of Denmark.

Video



Phil Chou, Francesco G.B. De Natale,
Enrico Magli, and Eckehard Steinbach

Trends in Multimedia Signal Processing

Pushed by the tremendous technological and societal changes of the last decade, multimedia is emerging more and more as the cornerstone of next generation information and communication technologies. Mobile devices, sensors, embedded systems, high-performance computing, broadband networks, and 3-D are some of the technologies that are pervading a generation of users that are more and more exigent and proactive. Evolving from a simple integration of existing technologies to a mutually aware, interdisciplinary development of new technologies is the biggest challenge of the multimedia SP (MMSP) community. In this article, we will list some of the most promising research trends that have recently emerged in MMSP.

TRENDS

Due to its intrinsic cross-disciplinary and highly dynamic nature, the area of multimedia is difficult to classify and structure. In the following sections, we will report some of the most interesting trends along three levels: multimedia systems (content versus architectures), multimedia delivery (content versus transport), and multimedia experience (content versus user).

MULTIMEDIA SYSTEMS

MULTIMEDIA/MULTIMODAL SYSTEMS

The possibility of spreading in the environment large numbers of sensors and actuators makes it possible to improve the perception capabilities of autonomous systems, but at the same time creates the need for intelligent strategies to handle huge amounts of information, taking care of the relevant redundancy, consistency, synchronization, and manipulation. Several problems are still open: where to do the processing (distributed versus centralized versus mixed solutions); how to integrate widely heterogeneous sources of information characterized by different reliability, dimension, and resolution; how to embed intelligence in the environment and in objects, providing lightweight platforms, operating systems, and algorithms; how to make those systems communicate with each other in an efficient, energy-aware, and secure way.

COGNITIVE/AWARE SYSTEMS

The major challenge is to go beyond ambient intelligence, creating systems that are able to understand, infer, and influence the environment: a big effort is needed to integrate extensive multisensory-multimodal systems (such as wireless sensor networks, multicamera, audio, radio-frequency identification, and body sensors). Some of the key trends in this field include: sensing (intelligent cooperative strategies for multisensorial integration, embedded intelligence, and distributed processing); beyond sensing (activity analysis, action recognition); beyond activities (individual behaviors, social behaviors, understanding roles and relationships); large/dense environments (outdoor, crowded areas, etc.); managing the complexity (power, bandwidth, computation). Application domains include homeland security, healthcare, assisted living (elderly, impaired), autonomous systems, and the Internet.

MULTIMEDIA DELIVERY

NEXT GENERATION NETWORKS VERSUS MEDIA DELIVERY NETWORKS

The development of new communication systems, in response to the increasing need of bandwidth, generates heterogeneity of coexisting networks and of content representations. This calls for the ability to interconnect such systems and formats, tackling problems such as network robustness, control protocols, security, quality of service (QoS)/experience, cross-layer optimization, and source coding. Multimedia streaming over such scenario poses significant technical challenges in the characterization of the streaming environment and the development of suitable network-aware streaming protocols. A specific case is given by cognitive radio systems, which exhibit a particularly dynamic behavior, posing unprecedented adaptation requirements for multimedia delivery. At the same time, many streaming systems employ content delivery networks. To become more effective, these networks must use information regarding the user preferences and current state to provide a personalized service. This includes collecting information about the user's context (e.g., position, speed, activity, and so on), and using this information to present to the user the content that he/she is more likely to be willing to consume at any given time.

MULTIMEDIA AND THE CLOUD

As cloud computing is becoming more and more available, the question raises of how to best exploit the possibility of offloading computations for multimedia applications. Thin client devices can certainly benefit from the cloud, which enables several traditionally impossible visual processing tasks. Research problems include how to partition visual processing tasks between cloud and client, how to solve the quality and adaptation problem of visual signals in the cloud, how to jointly optimize bandwidth, processing power, battery life, delay and QoE, how to enable offline experience through smart caching, how to exploit Web data available through the cloud, and how to provide security/privacy.

ADAPTIVE HTTP-BASED MEDIA DELIVERY

After almost 20 years of research into RTP over UDP/IP-based media delivery, practical system constraints (e.g., firewalls and caching solutions), the need for flexibility and simplicity, and design choices by large players (YouTube, Apple, etc.) have led to a rediscovery of TCP-based video

Digital Object Identifier 10.1109/MSP.2011.942320
Date of publication: 1 November 2011

delivery (both real time and on demand). In particular, http-based media delivery over mobile networks (e.g., LTE), the combination with scalable video and the question on how to adapt, schedule, and transmit the content in a receiver-driven, decentralized manner with the goal of maximizing the QoE have come back into focus both in industry and the MMSP research community.

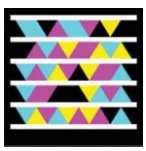
FUTURE OF PEER-TO-PEER

Recent years have seen a rapid increase of research on peer-to-peer media streaming systems. Current trends are targeting performance improvements of these systems in the key areas of reliability and security. The former aspect involves the design of efficient overlays and streaming protocols. Geographical information can be used to build and maintain overlays, reducing delays and failure probability; at the same time, technologies such as network coding enable new and more efficient streaming designs. Moreover, peer-to-peer systems have to face challenges that can hinder their widespread adoption, i.e., reputation, trust, security, and so on. Online social networks can come to the rescue, as social relationships can facilitate participation in peer-to-peer delivery and to contribute resources. How to leverage the best of both worlds for media delivery is a challenge.

SECURITY

Security is increasingly becoming an important area of multimedia SP. Classical security problems such as piracy prevention are now being complemented by security issues related to new multimedia applications. Such applications make extensive use of data provided by the users, but the users have no way to ensure that these data are not employed in ways that had not been authorized. An example is given by cloud computing, which is all about sending data to the cloud to have them processed remotely. Advances in secure multiparty computations are needed to face these new challenges, including the ability to perform processing directly on encrypted data.

Slides



MULTIMEDIA EXPERIENCE

NETWORKED MEDIA SEARCH AND BROWSING

Capturing user-versus-media semantics is still the big challenge: a strong integration is needed between media and knowledge communities to overcome the semantic gap. Some of the key trends in this field include representing and exploiting the context (the where, when, who, and what) to embed semantics into search; overcome the limits of taxonomies and ontologies (e.g., using social knowledge); design systems and techniques able to scale to extremely large archives (e.g., social media); capturing not only similarity but also diversity (the world's variety), facing computation/bandwidth/memory problems (enabling mobile search). Application domains include next generation media search, copy/duplication protection, augmented reality, geographic search, and event-based search.

QUALITY OF EXPERIENCE-DRIVEN MEDIA DELIVERY

For many years, QoS has been the focus for the analysis and optimization of media delivery systems. Both the characterization of networks in terms of their QoS support for media delivery as well as the QoS requirements of applications have driven research in this field. More recently, QoE has overtaken QoS as the metric of choice for the optimization of media delivery. Hot current research topics include objective QoE models for 2-D and 3-D video, QoE-driven X-layer optimization and resource allocation, large-scale QoE monitoring, and QoE prediction based on context information.

VIRTUAL/AUGMENTED REALITY AND TELEPRESENCE

There is a clear tendency that the real world and virtual worlds grow together and start overlapping. With the availability of real time and highly accurate information about a user's context, location and orientation, matching real observations with computer-generated information has become feasible and changes the way information is displayed. Compelling real and synthetically generated multimedia content plays a key role for new service generation and user acceptance. Similarly, our quest for immersive telepresence systems forces the MMSP community to look beyond the traditional audiovisual modalities. While the acquisition, processing, encoding, transmission, and display of digital audio and video has reached a high level of maturity, the processing and communication of haptic and physiological information has only recently received increasing attention. Along the same line, natural and self-explanatory user interfaces become more and more important. Application domains include immersive communication, telerobotics, telesurgery, telepresence, and remote monitoring and rehabilitation.

CONCLUSIONS

As a final remark, multimedia is a very active field, where different disciplines intersect creating great opportunities for new and stimulating research directions. Within the framework of multimedia, established SP technologies can evolve thanks to the "contamination" by other information and communication technologies fields (e.g., networking, sensors, systems, semantics) as well as other sciences (including social, psychological, physiological, and cognitive).

AUTHORS

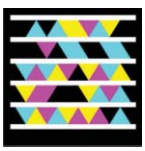
Phil Chou (pachou@microsoft.com) is a principal researcher at the Microsoft Research Center, Washington, United States.

Francesco G.B. De Natale (denatale@ing.unitn.it) is a professor in the Department of Information Engineering and Computer Science, University of Trento, Italy.

Enrico Magli (enrico.magli@polito.it) is a professor at the Dipartimento di Elettronica, Politecnico di Torino, Italy.

Eckehard Steinbach (eckehard.steinbach@tum.de) is a professor at LMT–Technical University of Munich, Germany. [SP]

Video



advertisers INDEX

The Advertisers Index contained in this issue is compiled as a service to our readers and advertisers: the publisher is not liable for errors or omissions although every effort is made to ensure its accuracy. Be sure to let our advertisers know you found them through *IEEE Signal Processing Magazine*.

ADVERTISER	PAGE	URL	PHONE
American Mathematical Society	175	www.ams.org/bookstore	
IEEE ICIP'12	5	www.icip2012.com	
IEEE Marketing Dept.	7	www.ieee.org/tryieeexplore	
Mathworks	CVR 4	www.mathworks.com/accelerate	+1 508 647 7040
Mini-Circuits	CVR 2, 3, CVR 3	www.minicircuits.com	+1 718 934 4500
University of Minnesota	177	www.ece.umn.edu	

advertising SALES OFFICES

James A. Vick
Staff Director, Advertising
Phone: +1 212 419 7767;
Fax: +1 212 419 7589
jv.ieeemedia@ieee.org

Marion Delaney
Advertising Sales Director
Phone: +1 415 863 4717;
Fax: +1 415 863 4717
md.ieeemedia@ieee.org

Susan E. Schneiderman
Business Development Manager
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org

Product Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, PA, DE, MD, DC, KY, WV

NEW ENGLAND/SOUTH CENTRAL/ EASTERN CANADA
Jody Estabrook
Phone: +1 774 283 4528;
Fax: +1 774 283 4527
je.ieeemedia@ieee.org
ME, VT, NH, MA, RI, CT, AR, LA, OK, TX
Canada: Quebec, Nova Scotia, Newfoundland, Prince Edward Island, New Brunswick

SOUTHEAST
Thomas Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
tf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

MIDWEST/CENTRAL CANADA
Dave Jones
Phone: +1 708 442 5633;
Fax: +1 708 442 7620
dj.ieeemedia@ieee.org
IL, IA, KS, MN, MO, NE, ND, SD, WI, OH
Canada: Manitoba, Saskatchewan, Alberta

MIDWEST/ ONTARIO, CANADA
Will Hamilton
Phone: +1 269 381 2156;
Fax: +1 269 381 2556
wh.ieeemedia@ieee.org
IN, MI. Canada: Ontario

WEST COAST/MOUNTAIN STATES/ WESTERN CANADA
Marshall Rubin
Phone: +1 818 888 2407;
Fax: +1 818 888 4907
mr.ieeemedia@ieee.org
AZ, CO, HI, NM, NV, UT, AK, ID, MT, WY, OR, WA, CA. Canada: British Columbia

EUROPE/AFRICA/MIDDLE EAST
Heleen Vodegel
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
hv.ieeemedia@ieee.org
Europe, Africa, Middle East

ASIA/FAR EAST/PACIFIC RIM
Susan Schneiderman
Phone: +1 732 562 3946;
Fax: +1 732 981 1855
ss.ieeemedia@ieee.org
Asia, Far East, Pacific Rim, Australia, New Zealand

Recruitment Advertising

MIDATLANTIC
Lisa Rinaldo
Phone: +1 732 772 0160;
Fax: +1 732 772 0164
lr.ieeemedia@ieee.org
NY, NJ, CT, PA, DE, MD, DC, KY, WV

NEW ENGLAND/EASTERN CANADA
Liza Reich
Phone: +1 212 419 7578;
Fax: +1 212 419 7589
e.reich@ieee.org
ME, VT, NH, MA, RI. Canada: Quebec, Nova Scotia, Prince Edward Island, Newfoundland, New Brunswick

SOUTHEAST
Cathy Flynn
Phone: +1 770 645 2944;
Fax: +1 770 993 4423
cf.ieeemedia@ieee.org
VA, NC, SC, GA, FL, AL, MS, TN

MIDWEST/SOUTH CENTRAL/ CENTRAL CANADA
Darcy Giovingo
Phone: +1 847 498 4520;
Fax: +1 847 498 5911
dg.ieeemedia@ieee.org;
AR, IL, IN, IA, KS, LA, MI, MN, MO, NE, ND, SD, OH, OK, TX, WI. Canada: Ontario, Manitoba, Saskatchewan, Alberta

WEST COAST/SOUTHWEST/ MOUNTAIN STATES/ASIA
Tim Matteson
Phone: +1 310 836 4064;
Fax: +1 310 836 4067
tm.ieeemedia@ieee.org
AZ, CO, HI, NV, NM, UT, CA, AK, ID, MT, WY, OR, WA. Canada: British Columbia

EUROPE/AFRICA/MIDDLE EAST
Heleen Vodegel
Phone: +44 1875 825 700;
Fax: +44 1875 825 701
hv.ieeemedia@ieee.org
Europe, Africa, Middle East

Digital Object Identifier 10.1109/MSP.2011.941850

[in the SPOTLIGHT]

"Trends" Expert Overview Sessions Revived at ICASSP 2011: Part 2

INTRODUCTION

This is the second in a series of three columns summarizing the "Trends" expert sessions organized by the Signal Processing Society Technical Committees during ICASSP 2011 in Prague, Czech Republic. Readers have an opportunity to access these Trends session summaries authored by the Technical Committees.

Alle-Jan van der Veen and Jose C. Principe

—Note: Additional multimedia resources for these sessions can be found at http://www.signalprocessingsociety.org/publications/periodicals/spm/columns-resources/. Alternatively, the Tag to the left can be scanned using your smart phone (with a free reader download) to access this Web site.

Several other Tags in this article can be scanned to access video or slides associated with the article.

Digital Object Identifier 10.1109/MSP.2011.942869
Date of publication: 1 November 2011

Trends in Bioimaging and Signal Processing

Jean-Christophe Olivo-Marin, Michael Unser, Laure Blanc-Féraud, Andrew Laine, and Boudewijn Lelieveldt

The area of bioimaging and signal processing (BISP) is concerned with the development of dedicated tools for biomedical and bioinformatics applications and with fostering interdisciplinary and integrative approaches of biomedical topics by bridging the biology, medicine, and SP communities [1]. Increasingly sophisticated imaging devices and protocols are now being developed for biological and medical imaging. Understanding functional and pathological mechanisms in patients or model organisms indeed requires being able to visualize and measure in vivo and in situ at scales ranging from subcellular to whole body. A similar trend occurs in the area of genomics and proteomics, where sophisticated bioinformatics methods are required to assemble, decipher, and compare data from large-scale sequencing data [2]. This article focuses on the latest developments in biomedical imaging that has sparked a full array of challenging topics for the signal and image processing community.

Digital Object Identifier 10.1109/MSP.2011.942317
Date of publication: 1 November 2011

IMAGE PROCESSING CHALLENGES IN BIOIMAGING

With the recent development of fluorescent probes and of new high-resolution microscopes [e.g., confocal, two-photon, stimulated emission depletion (STED), photo activated light microscopy (PALM), stochastic optical reconstruction microscopy (STORM)], biological imaging has grown quite sophisticated and is presently having a profound impact on the way research is being conducted in cell biology [3]. Biomedical scientists can visualize subcellular components and processes, both structurally and functionally, in two or three dimensions, at different wavelengths (spectroscopy), and they can perform time-lapse imaging to investigate cellular dynamics. Researchers are faced with an ever-increasing volume of data to visualize, analyze, and process; in particular, in the context of high throughput screening that requires the engineering of fast, robust algorithms with minimal user interaction. We have grouped the related SP challenges in three broad categories:

- *Image reconstruction and mathematical imaging*: While there have been tremendous advances in the instrumentation, it is very likely that the capabilities of modern microscopes can be further enhanced with the help of sophisticated SP for denoising, three-dimensional (3-D) deconvolution, and/or tomographic reconstruction. The problems that are specific to this kind of imaging are high levels of noise (photon-counting statistics), photo-bleaching, and the presence of aberrations (in particular, the depth dependence of the point spread function). In addition, the challenge is to design true 3-D (or 3-D + time) reconstruction algorithms that are fast enough to be used in practice. The foundation of such methods is a

(continued on page 191)

1053-5888/11/$26.00©2011IEEE